Exploring the genesis and specificity of serum antibody binding

Mathematical modeling and data analysis of antibody-peptide reactivity data

D I S S E R T A T I O N

zur Erlangung des akademischen Grades

Dr. rer. nat.
im Fach Biologie

eingereicht an der
Mathematisch-Naturwissenschaftlichen Fakultät I
Humboldt-Universität zu Berlin

von
Dipl.-Biol. Victor Greiff

Präsident der Humboldt-Universität zu Berlin:
Prof. Dr. Jan-Hendrik Olbertz

Dekan der Mathematisch-Naturwissenschaftlichen Fakultät I:
Prof. Stefan Hecht PhD

Gutachter:
1. Dr. Michal Or-Guil
2. Prof. Dr. Edda Klipp
3. Prof. Dr. Birgit Sawitzki

eingereicht am: 31.07.2012
Tag der mündlichen Prüfung: 13.12.2012

»Und wenn ich dann Kunde von Heilmann und Coppi erhielte, würde meine Hand auf dem Papier lahm werden. Ich würde mich vor den Fries begeben, auf dem die Söhne und Töchter der Erde sich gegen die Gewalten erhoben, die ihnen immer wieder nehmen wollten, was sie sich erkämpft hatten, Coppis Eltern und meine Eltern würde ich sehn im Geröll, es würde pfeifen und dröhnen von den Fabriken, Werften und Bergwerken, Tresortüren würden schlagen, Gefängnistüren poltern, ein immerwährendes Lärmen von eisenbeschlagnen Stiefeln würde um sie sein, ein Knattern von Salven aus Maschinenpistolen, Steine würden durch die Luft fliegen, Feuer und Blut würden aufschießen, bärtige Gesichter, zerfurchte Gesichter, mit kleinen Lampen über der Stirn, schwarze Gesichter, mit glitzernden Zähnen, gelbliche Gesichter unterm Helm aus geflochtnem Bast, junge Gesichter, fast kindlich noch, würden anstürmen und wieder untertauchen im Dampf, und blind geworden vom langen Kampf würden sie, die sich auflehnten nach oben, auch herfallen übereinander, einander würgen und zerstampfen, wie sie oben, die schweren Waffen schleppend, einander überrollten und zerfleischten, und Heilmann würde Rimbaud zitieren, und Coppi das Manifest sprechen, und ein Platz im Gemenge würde frei sein, die Löwenpranke würde dort hängen, greifbar für jeden und solange sie unten nicht abließen voneinander, würden sie die Pranke des Löwenfells nicht sehn, und es würde kein Kenntlicher kommen, den leeren Platz zu füllen, sie müßten selber mächtig werden dieses einzigen Griffs, dieser weit ausholenden und schwingenden Bewegung, mit der sie den furchtbaren Druck, der auf ihnen lastete, endlich hinwegfegen könnten.«

— Peter Weiss, Die Ästhetik des Widerstands

Meiner Familie

Abstract

Human and murine humoral immune responses are associated with changes of both the composition and the concentration of serum antibodies. Signal intensity-based antibody binding profiles measured with random-sequence peptide microarrays attempt to capture these changes to render them applicable to serological diagnostics. Diagnostics based on antibody profiling rest primarily on the assumption that profiles of diseased and healthy individuals differ consistently from one another. The challenge for antibody profiling lies in reflecting the change in the antibody mixture induced by the disease while taking into account the variability of antibody profiles of healthy individuals. In this work, the antibody repertoire's impact on antibody binding profiles is studied. Since the characterizing components of polyclonal antibody mixtures, such as composition and concentration, are difficult to study in vitro a mathematical model for antibody-peptide binding was formulated.

This model is based on the law of mass action and incorporates as parameters (i) antibody and peptide sequences and (ii) antibody concentrations. The binding affinity of simulated monoclonal antibodies depends *non-linearly* on amino acid positions in the peptide sequences. The model was both mathematically analyzed and implemented in silico to simulate antibody-peptide binding data. Mathematical analysis and simulations predicted that mixtures of highly diverse random antibodies which are not dominated concentration-wise by few antibodies—termed unbiased mixtures—could be *linearly* predicted based *only* on the amino acid composition of the peptide library used. Thus, any unbiased mixture independent of its specific antibody composition yields the same antibody binding profile for a given peptide library. This linear relationship led to the formulation of a linear regression model of which amino acid associated-weights (AAWS) emerge as near perfect predictors of antibody binding profiles. AAWS indicate the contribution of every amino acid to signal intensity and are a compact, lossless representation of unbiased mixtures' antibody binding profiles. For low-diversity antibody mixtures, this linear regression model breaks down.

In order to test the in vitro relevance of the mathematically predicted ensemble properties of antibody mixtures, monoclonal (low antibody diversity) and serum antibodies (high antibody diversity) were incubated with the same peptide library. Indeed, as predicted by theory, the predictive performance of AAWS was significantly higher for antibody binding profiles of serum than of monoclonal antibodies. In addition, AAWS, and to a lesser extent antibody binding profiles, were found to be consistent across healthy individuals, both murine and human, thereby showing the independence of antibody binding profiles and AAWS on the specific antibody mixture. The concept of unbiased mixtures best approximates sera of healthy individuals.

Simulated antibody binding profiles of mixtures biased by random dominant antibodies were found to be isotropically distributed in the variance space. Consequently, to separate simulated antibody binding profiles into different groups, antibody-peptide binding of dominant antibodies had to be consistent across individuals of a given group but different from any other. The intra-group consistency of antibody-peptide binding is a basic premise of serological diagnostics: the mathematical model does not only fulfill this premise, but also predicts antibody dominance as a condition which is able to establish classifiable intra-group consistency.

In particular, antibody dominance caused—unlike variations in total antibody concentration—rank changes in the simulated antibody binding profiles. In vitro, rank changes were consistent across healthy and diseased mice thus serving to classify mice by stage of immune response. Additionally, ranks of antibody binding profiles of plasma samples from healthy volunteers obtained over the course of one month clustered by volunteer. This indicates the need for serological methods to take into account individual variability to detect disease-induced changes in antibody mixtures.

Furthermore, simulations showed that AAWS are highly noise-resistant: AAWS could readily separate original signal intensities from noise over a large range of noise amplitudes. In fact, AAWS were found to not only show high consistency across sera incubated on the same batch, but, unlike antibody binding profiles, also across batches. However, AAWS varied with the microarray manufacturer.

In conclusion, this work shows that serum antibody ensemble properties impact the genesis of antibody binding profiles measured with random-sequence peptide microarrays. This thesis indicates that a knowledge of both a polyclonal mixture's diversity and composition is essential for the interpretation of antibody binding profiles with respect to both serological diagnostics and B-cell epitope mapping. Specificity, and thus classifiability, of serum antibody binding profiles is a function of both the investigated antibody mixtures and technological features.

Zusammenfassung

Menschliche und murine humorale Immunantworten gehen einher mit der Veränderung der Zusammensetzung und der Konzentration von Serumantikörpern. Signalintensitäts-basierte Antikörperbindungsprofile, gemessen mit Zufallspeptidmikroarrays, versuchen diese Veränderungen zu detektieren, um sie für serologische Diagnostik nutzbar zu machen.

Die auf Antikörper-Profiling basierende Diagnostik beruht auf der Annahme, dass Antikörperbindungsprofile von kranken und gesunden Individuen sich systematisch voneinander unterscheiden. Antikörper-Profiling muss sowohl die krankheitsinduzierte Veränderung der Antikörpermischung wiedergeben als auch der Variabilität Antikörperprofilen gesunder Individuen Rechnung tragen.

Gegenstand dieser Arbeit ist die Analyse des Einflusses des Antikörperrepertoires auf Antikörperbindungsprofile. Da die charakteristischen Komponenten polyklonaler Antikörpermischungen, wie Zusammensetzung und Konzentration, experimentell überwiegend nicht quantifizierbar sind, wurde ein mathematisches Modell für Antikörper-Peptidbindung aufgestellt.

Dieses Modell basiert auf dem Massenwirkungsgesetz und beinhaltet als Parameter (i) Antikörper- und Peptidsequenzen sowie (ii) Antikörperkonzentrationen. Die Bindungsaffinität simulierter monoklonaler Antikörper hängt *nichtlinear* von den Aminosäurepositionen in den Peptidsequenzen ab. Das Modell wurde mathematisch analysiert und in silico implementiert, um Antikörperbindungsprofile zu simulieren. Mathematische Analyse und Simulationen ergaben, dass die Antikörperbindungsprofile von Mischungen hochdiverser, zufällig generierter Antikörpersequenzen, welche nicht durch wenige Antikörper konzentrationsdominiert sind – genannt ideale Mischungen – *linear ausschließlich* mit Hilfe der Aminosäurezusammensetzung der Peptidbibliothek vorhergesagt werden können. Das bedeutet, dass für eine gegebene Peptidbibliothek alle idealen Mischungen unabhängig von ihrer Zusammensetzung das gleiche Antikörperbindungsprofil erzielen. Dieser lineare Zusammenhang führte zu der Formulierung eines linearen Regressionsmodells, aus welchem Aminosäure-assoziierte Gewichte (AAWS) hervorgehen. AAWS sind fast perfekte Prädiktoren von Profilen idealer Mischungen. Die AAWS geben den Anteil jeder Aminosäure zur gemessenen bzw. simulierten Peptid-Signalintensität wieder. Sie stellen eine kompakte, verlustfreie Abbildung von Antikörperbindungsprofilen idealer Mischungen dar. Für niedrig-diverse Mischungen ist die Vorhersagekraft des Regressionsmodells jedoch eingeschränkt.

Um die in vitro-Relevanz der mathematisch vorhergesagten Ensembleeigenschaften von Antikörpermischungen zu überprüfen, wurden monoklonale Antikörper (niedrige Antikörperdiversität) und Serumantikörper (hohe Antikörperdiversität) mit derselben Peptidbibliothek inkubiert. Wie durch das Modell vorhergesagt war (i) die AAWS-Vorhersagekraft signifikant höher für Antikörperbindungsprofile von Serum- als für monoklonale Antikörper. (ii) Des Weiteren entsprachen sich die AAWS gesunder Individuen (murin, human) und in einem geringeren Maße auch deren Antikörperbindungsprofile. In der Tat sind AAWS in gewissem Umfang von der spezifischen Antikörperkomposition unabhängig. Das Konzept der idealen Mischung entspricht bevorzugt Seren gesunder Individuen.

Simulierte Antikörperbindungsprofile von zufällig dominierten Antikörpermischungen waren isotrop im Varianzraum verteilt. Folgerichtig konnten in Simulationen

nur solche Antikörperbindungsprofile dominierter Antikörpermischungen in Gruppen aufgetrennt werden, deren Antikörperpeptidbindung sich innerhalb einer Gruppe ähnelte und sich von jeder anderen Gruppe unterschied. Die Intra-Gruppen-Konsistenz der Antikörperpeptidbindung ist einer der Hauptprämissen serologischer Diagnostik. Das mathematische Modell erfüllt diese nicht nur, sondern hebt prädiktiv Antikörperdominanz als einen Zustand hervor, der Intra-Gruppen-Konsistenz herbeizuführen vermag.

Im Gegensatz zu Gesamtantikörperkonzentrationsschwankungen führte insbesondere Antikörperdominanz zu Peptid-Rangveränderungen innerhalb der simulierten Antikörperbindungsprofile. In vitro konnten konsistente Rangunterschiede festgestellt werden, welche die Klassifizierung von Seren gesunder und parasiteninfizierter Mäuse ermöglichten. Außerdem clusterten die Ränge der Antikörperbindungsprofile von Plasmaproben, erhalten über einen Zeitraum von einem Monat von gesunden Menschen, bezüglich des jeweiligen gesunden Individuums. Dies belegt, dass serologische Methoden individuelle Variabilität in Betracht ziehen müssen, um krankheitsinduzierte Veränderungen diagnostizieren zu können.

Weiterhin zeigten Simulationen, dass AAWS hochgradig rauschresistent sind. AAWS konnten das simulierte Originalsignal vom verrauschten Signal über eine große Bandbreite von Rauschamplituden trennen. Darüber hinaus waren AAWS nicht nur serum-, sondern im Gegensatz zu Antikörperbindungsprofilen, auch produktionschargenunabhängig. Jedoch hingen die AAWS vom Mikroarray-Produzenten ab.

Zusammenfassend zeigt diese Arbeit, dass Antikörper-Ensembleeigenschaften die Genese von mit Zufallspeptidemikroarrays bestimmten Antikörperbindungsprofilen beeinflussen. Kenntnisse über die Zusammensetzung einer polyklonalen Mischung sind essentiell für die Interpretation von Antikörperbindungsprofilen in Bezug auf serologische Diagnostik und Epitopkartierung. Die Spezifität und damit auch die Klassifizierbarkeit von Antikörperbindungsprofilen ist sowohl eine Funktion der untersuchten Antikörpermischung als auch technologischer Faktoren.

Danksagung

»If a machine is expected to be infallible, it cannot also be intelligent.«
— Alan Mathison Turing

Diese Doktorarbeit entstand am Institut für Biologie der Humboldt-Universität zu Berlin unter der Leitung von Frau Dr. Michal Or-Guil.
Außerordentlicher Dank gilt:

- Atijeh Valai für ihre in jeglicher Hinsicht kompetente Unterstützung,
- Henning Redestig für die wegweisende Hilfestellung beim Fertigstellen des BMC-Artikels,
- Johannes Eckstein als Ansprechpartner bei physikalischen Fragen,
- René Riedel für die Endredaktion von Manuskripten und der Doktorarbeit,
- Carsten Mahrenholz für fachlichen und persönlichen Rat,
- Ulrich Bodenhofer und Sepp Hochreiter für Beantwortung von Fragen bezüglich P-SVM,
- und Johannes Schuchhardt für das Vermitteln entscheidender Einsichten in Bezug auf das mathematische Modell.

Ich möchte weiterhin die Unterstützung von Juliane Lück, Armin Weiser, Christin Schläwicke, Stefan Kröger, Bodo Steckel, Ata Valai, Clarissa Wild, Katja Köhler, Harald Seitz, Matthias Kröger, Nicole Wittenbrink und Sebastian Rausch bei der Bearbeitung von Teilabschnitten dieser Arbeit hervorheben.

Außerdem danke ich den ehemaligen Mitgliedern der AG Systemimmunologie für ihre Unterstützung: André Dautcourt, Ludwig Weh, und Nicole Bruni.

Weiteren Menschen, denen ich zu Dank verpflichtet bin: der AG Hamann im RCIS (Francesca Liu, Jennifer Pfeil, Elisabeth Kenngott, Ute Hoffmann, Uta Lauer), Ilko Kastirr, Svenja Steinfelder, Jana Krietsch, Chris Bauer, Susanne Hartmann, Hedda Wardemann, Rafael Burtet und Andrea Maranhao.

Ich danke Frau Dr. Michal Or-Guil für die Ermöglichung der Anfertigung meiner Dissertationsschrift in ihrer Arbeitsgruppe.

Contents

List of Figures

List of Tables

List of Abbreviations

AACM Amino acid composition matrix

AAWS Amino acid-associated weights (assigned: $\vec{h}$, estimated: $\vec{w}$)

Ab Antibody

Ag Antigen

AP Acute phase

AR Antibody repertoire

ASC Antibody secreting cell

BACC Balanced accuracy

BCR B-cell receptor

BSA Bovine serum albumin

C Constant

CDR Complementarity determining region

CLT Central limit theorem

CP Early chronic phase

CR Chronic rejection

dpi Days post-infection

EIA Enzyme immunoassay

Fab Fragment antigen binding

Fc Fragment crystallizable

GC Germinal center

HB *Heligmosomoides bakeri*

HE Healthy

HRP Horseradish peroxidase

i.i.d. independent and identically distributed

Ig Immunoglobulin

κ Kappa

λ Lambda

LOOCV Leave-one-out cross-validation

mAb Monoclonal antibody

MS Mouse study

NS NOD study

OLR Ordinary least squares regression

OT Operational tolerant

PBS Phosphate buffered saline

PC Physico-chemical

PCA Principal component analysis

PLSR Partial least squares regression

P-SVM Potential support vector machine

SHM Somatic hypermutation

SHS Slovenian healthy study

SI Signal intensity

SNC Self-non-self criterion

ST Stable

TB Tuberculosis

TD T-cell dependent

TI T-cell independent

V Variable

VSV Vesicular stomatitis virus

1 Introduction

1.1 The mammalian immune system

The mammalian immune system[1] is responsible for removing dead and non-functioning cells [4, 5] as well as for clearing the body from xenobiotics and pathogens such as bacteria, viruses, fungi and protozoa. A complex system of immune cells is distributed throughout the body. In addition, several immune organs exist. In the primary organs, bone marrow and thymus, immune cells are generated and mature, whereas in secondary tissues, including spleen and lymph nodes, the processing of the immune response takes place[2].

The immune system relies on three pillars: (i) physico-chemico-mechanical immune barriers [7], (ii) the innate immune system and (iii) the adaptive immune system[3].

After pathogens have passed the physical immune barriers, they are first challenged by the innate immune system. It comprises germline encoded immune mediators such as cytokines and complement as well as immune cells (e.g. macrophages, granulocytes and NK cells). The innate immune system causes local inflammation, which preludes the active and rapid elimination of pathogens by either phagocytosis or cell lysis. Additionally, the innate immune system stimulates the adaptive immune system, which induces both a highly diverse and specific immune response, and immunological memory. Immunological memory is defined as a concept, which enables the immune system to react more specific, faster and with higher amplitude to already encountered pathogens [2].

Adaptive immunity essentially branches off in two arms. The cell-mediated immunity relies on cytotoxic T lymphocytes (T cells) that are responsible for the elimination of intracellular pathogens by either destroying them or by lysing infected cells. On the contrary, the humoral immunity is mainly mediated by glycoproteins called antibodies that are derived from B lymphocytes (B cells). Antibodies account for the major defense against extracellular[4] pathogens and their toxins. Antibody binding neutralizes the targets, marks them for elimination (opsonization) and activates adequate effector mechanisms. Effector mechanisms comprise among others activation of complement and endocytosis by antigen presenting cells.

[1] In the following, I will focus mainly on the human and murine immune system the both of which are similar in some respects but different in others [1]. Parts of this introduction are inspired by the textbooks *Immunobiology* [2] and *Cellular and molecular immunology* [3].

[2] The existence of "tertiary" (or ectopic) lymphoid organs was also reported. They are characterized as cellular accumulations arising during chronic inflammation by the process of lymphoid neogenesis [6].

[3] Newer studies draw a rather interconnected picture of innate and adaptive immune system in which both parts critically depend on one another [8].

[4] The existence of intracellular antibody-mediated immunity has recently been suggested [9].

1.2 Humoral immunity

1.2.1 Immune reaction and immune response

Following Thomas Pradeu, in this work an immune reaction is defined as the biochemical interaction between an immune receptor and its ligand. An immune response is launched only if its immunological effector mechanisms were activated [10].

1.2.2 Antigens and immunogenicity

An antigen is defined as any substance that can bind to a specific antibody [2]. The antigen's ability to induce an immune response in a competent host is known as immunogenicity. The term immunogenicity has no meaning outside the host context and depends on the potentialities of the host being immunized such as its immunoglobulin gene repertoire and various cellular regulatory mechanisms [11, 12].

1.2.3 Criteria for immunogenicity

Since the 1950s, a consensus has formed on the acceptance, and the adjustment of Burnet's seminal ideas [13–15] according to which the discrimination between "self" and "non-self" is the criterion (SNC) for immunogenicity [16]: every element that distinctively belongs to the organism ("self") does not trigger an immune response, whereas every foreign element ("non-self") triggers an immune response [17]. Yet, several published experimental data [5, 18–20] as well as conceptual articles [21, 22] have put the SNC into question.

At least two other significant alternatives to the SNC have been proposed. (i) Polly Matzinger formulated the danger theory, wherein the immune system does not react to non-self but rather to any danger, be it exogenous or endogenous [19, 23]. (ii) The continuity criterion, published by Thomas Pradeu and Edgardo D. Carosella, relies on the immune system's ability to discriminate pathogens based on significant molecular differences. Thus, the immune system does not respond to non-self, but rather to abrupt modifications of the antigenic patterns with which it is in contact [15].

1.3 The antibody molecule

1.3.1 The function of antibody molecules

The antigen recognition molecules of B cells are the immunoglobulins (Igs, Figure 1.1). These proteins are produced by B cells in a vast variety, each B-cell clone producing an Ig of a single kind. Membrane-bound Ig on the B-cell surface serves as the cell's receptor for antigen, and is known as the B-cell receptor (BCR). Igs of the same antigen affinity are secreted as antibodies by antibody secreting cells (ASCs)—proliferating plasmablasts and terminally differentiated plasma cells.

The antibody molecule has two distinct roles: (i) binding to molecules associated with the immune response eliciting pathogen in a neutralizing fashion and (ii) recruiting

additional cells (and molecules) to the site of inflammation in order to destroy opsonized pathogens.

The twofold functionality of antibodies is also mirrored by the structural duality of the antibody. One part of the antibody recognizes and binds to the antigen, whereas the other one engages different effector functions. The antigen-binding region varies extensively among antibody molecules and is thus named V(ariable) region (Section 1.3.3). The part of the antibody which engages effector functions does not vary in the same way and is thus called the C(onstant) region. It is generated in five main forms, which are specialized for activating different effector functions[5].

The membrane-bound BCR does not have these effector functions, because the C region remains inserted in the membrane of the B cell. Its function is as a receptor that recognizes and binds antigen thereby transmitting signals, which elicit mechanisms such as (T-cell mediated) B-cell activation, clonal expansion and the production of antibodies [3].

1.3.2 The structure of antibody molecules

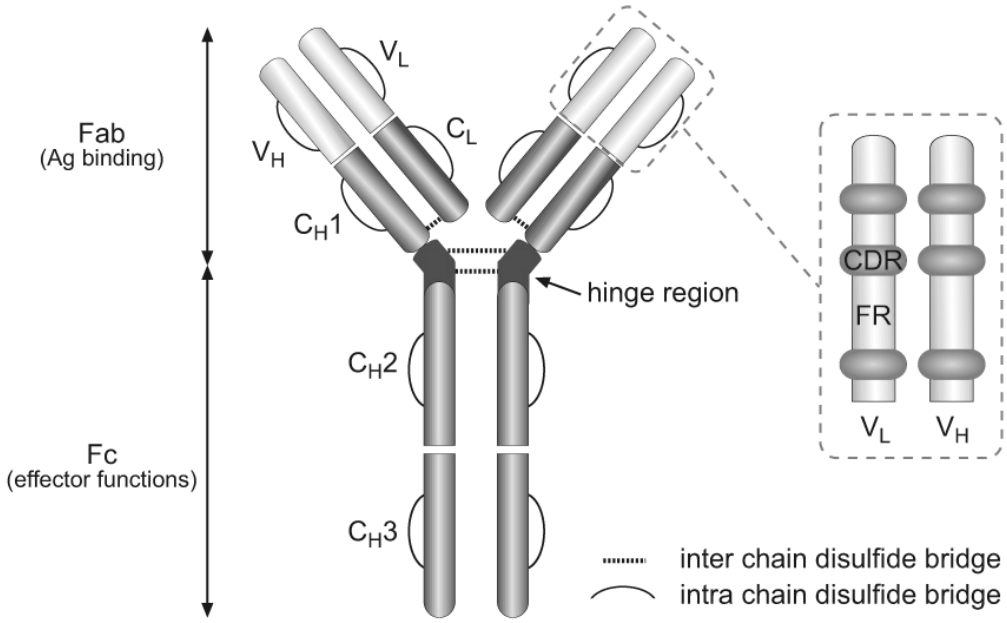


Figure 1.1: Schematic depiction of the IgG molecule. The antigen binding sites are formed by juxtaposition of variable light chain (V_L) and variable region heavy chain domains (V_H). C: constant region, CDR: complementarity determining region, Fab: fragment antigen binding, Fc: fragment crystallizable, FR: framework, H: heavy chain, L: light chain, V: variable region. From Wittenbrink (PhD thesis, [27]) who modified this Figure from Abbas and Lichtman [3].

Antibodies are roughly Y-shaped molecules consisting of three equal-sized portions connected by disulfide bonds (Figure 1.1). The five main Ig classes (also called isotypes)

[5]The C region can affect the interaction of the V region with an antigen [24–26].

IgA, IgD, IgE, IgM[6], IgG are mainly distinguished by their C region. In the following, the IgG molecule is described in more detail exemplifying the general structure of Igs as it is the most abundant isotype (Section 1.6). Nevertheless, the general structural features of all Ig isotypes are similar.

IgGs are large proteins of about 150 kDa consisting of two kinds of polypeptide chains. The one with a molecular weight of 50 kDa is referred to as H chain and the other of 25 kDa is called L chain. Each IgG molecule consists of two heavy and two light chains. The two H chains are linked together by disulfide bonds, and each H chain is linked to an L chain by another disulfide bond. In any given Ig molecule, these chains are identical enabling the Ig to bind simultaneously to two identical structures.

Light chains are subdivided into kappa (κ) and lambda (λ) chains. Each antibody has either κ or λ light chains, never both together. The class, and thus the effector function of an antibody is defined by the structure of its heavy chain. The distinctive function of the several classes results from the properties conferred to them by the carboxyl terminal part of the H chain, where it is not associated with a light chain.

1.3.3 Variability of antibody molecules

Each Ig chain consists of similar, although not identical about 100 amino acid long sequences. Each of these repeats corresponds to a discrete, compactly folded region of protein structure known as a protein domain. The light chain is made up of two of such Ig domains, whereas the heavy chains of the IgG molecule consists of four of such domains. The aminoterminal sequences of heavy and light chain vary markedly among antibodies. The sequence variability is limited to the first 110 amino acids, corresponding to the first domain, whereas the remaining domains are constant between Ig molecules of the same isotype. The amino-terminal V domain of the heavy and light chains (V_H and V_L respectively) together make up the V region of the molecule and confer on it the ability to specifically bind antigen, whereas the C domain of the heavy and light chains (C_H and C_L respectively) make up the C region of the heavy and light chains. In order to dissect the function of the parts of the antibody, proteases have been used, cleaving the antibody in distinct polypeptide sequences. Papain (a protease) cleaves the antibody into three fragments. Two fragments are identical and contain the antigen-binding activity. These are termed Fab[7] fragments. The remaining fragment shows no antigen-binding activity but is crystallizable, thus termed Fc fragment[8]. It represents the part of the antibody that interacts with effector molecules and cells. The reasons for effector-functional differences between H chain-isotypes lie mainly in the Fc fragment.

Each B-cell clone produces antibodies with a unique V region. The V region's sequence variability is concentrated in three hypervariable segments denoted as HV1, HV2, and HV3. They are found in both the V_H and V_L regions.

The most variable part of the domain is the HV3 region. The less variable regions between the hypervariable regions, which comprise the rest of the V domain are termed

[6] IgM is the only isotype common to all vertebrates [28].
[7] Fragment antigen binding.
[8] Fragment crystallizable.

framework regions. Four of such regions exist in each V domain, termed FR1 to FR4.

When the V_L and V_H domains are paired in the antibody molecule, the hypervariable regions from each domain are brought together, creating a single hypervariable site at the top of each arm of the molecule. These are the sites mostly involved in antigen-binding. The six hypervariable regions determine antigen affinity by forming a surface complementary to the antigen and are more commonly termed complementarity determining regions (CDRs) denoted CDR1 to CDR3 (there are three CDRs from each of the heavy and light chains).

1.4 Antibody reactivity

1.4.1 B-cell epitopes

B-cell epitopes are traditionally defined as antigenic molecules that are recognized by individual antibody paratopes: the epitope is the molecular surface that makes physical contact with the paratope [29, 30]. Greenspan and van Regenmortel suggest an operational epitope definition according to which epitope and paratope are relational entities defined by their mutual complementarity. An epitope is thus a function or an activity, as opposed to a mere structure [12, 31, 32].

Epitopes are usually classified as either continuous (obsolete: linear) or discontinuous (obsolete: conformational). Epitopes on the surface of proteins are mostly discontinuous and conformation-dependent [33–35]. The label *continuous epitope* is given to any short, linear peptide fragment of the antigen that binds to antibodies raised against the intact protein. Because the peptide fragment usually does not retain the conformation present in the folded protein and mostly represents only a portion of a more complex epitope, it tends to react only weakly with anti-protein antibodies. Discontinuous epitopes are made up of residues brought together by the folding of the polypeptide chain. Thus, as a rule, antibodies to discontinuous epitopes will recognize the antigen only if the protein molecule is intact, and its native conformation is preserved. There are exceptions, however, and it has been estimated that about 10% of the monoclonal antibodies that recognize discontinuous epitopes are also able to react with linear peptide fragments of the protein [12].

It is now accepted that the entire surface of the protein harbors numerous overlapping epitopes [34]. For example, insulin, a dimeric protein with 51 amino acids, has on its surface at least 115 B-cell epitopes [12, 36].

1.4.2 Antibody-epitope interaction

The central paradigm of antigen-antibody recognition is that the three-dimensional structure formed by the six CDRs recognizes and binds a complementary surface (epitope) on the antigen (Sections 1.3.3 and 1.4.1, [37][9]).

[9] However, Kunik and colleagues very recently showed that about 20% of the amino acid residues that bind the antigen fall outside the CDRs [38].

An antibody contains a variety of binding sites. Each antibody binding site defines a paratope composed of the particular amino acids of that antibody that physically bind to a specific epitope. Approximately 50 variable amino acids make up the potential binding area of an antibody [32]. Typically, only about 15 of these 50 amino acids physically contact a particular epitope. These 15 contact residues define the structural paratope. Only approximately 5 of these amino acids dominate in terms of binding energy. In both epitope and paratope, substitutions both in and away from the binding site can change the spatial conformation of the binding region and affect the binding reaction [32, 39, 40].

The association of antibody and antigen is of non-covalent nature. The free energy of interaction between an antibody and its antigen is a function of both enthalpy and entropy. Non-bonded forces between the interacting molecules include hydrophobic, hydrogen bonds, van der Waals and electrostatic interactions [41]. Charge neutralization in the interface plays a prominent role as well [42].

CDRs were found to have a much greater frequency of tyrosine and tryptophan residues than is usual on the surface of protein molecules [33, 43–45]. These aromatic side chains can make large rotations with little entropic cost, and they contribute significantly to the binding energy [12, 41]. Furthermore, crystallographic studies showed that binding involved a certain amount of induced fit [12, 46]. Upon binding, residues are displaced by several angstrom [47, 48].

Furthermore, two molecules that have nearly identical structures on the basis of crystallography may not interact comparably with a given receptor because of differences in molecular dynamics [49]: the crystallographic structure of antibody-antigen complex captures merely one point in time. The contributions of the time dimension should therefore be taken into account for a characterization of bimolecular interactions [31, 32].

Hence, Greenspan proposed a richer epitope description by taking into account (i) the spatial coordinates of the contact atoms, (ii) the dynamics [time dimension] of the atoms involved in contact with the paratope, (iii) the relative energetic contributions of atoms or amino acids to the interaction or to the discrimination between cognate epitopes and other epitopes as well as (iv) the context in which the binding takes place [31, 50].

1.4.3 Affinity and avidity

The affinity between a ligand (such as an antibody) and a protein (such as an antigen) (Equation 1.1), defining the strength of a ligand-protein bond, is commonly expressed by the dissociation constant K_d (Equation 1.2).

$$\mathrm{Ab} + \mathrm{Ag} \rightleftharpoons \mathrm{AbAg} \tag{1.1}$$

$$K_\mathrm{d} = \frac{[\mathrm{AbAg}]}{[\mathrm{Ab}]\,[\mathrm{Ag}]} = \frac{1}{K_\mathrm{a}} \tag{1.2}$$

In the specific case of antibodies binding to antigen, usually the affinity constant, defined at chemical equilibrium, K_a, is used (Equation 1.2). It is the inverted dissociation constant and determines the binding strength of an antibody with a given antigen. The

higher the affinity of the antibody for its antigen, the less antibody is required to eliminate the antigen in a physiological immune response, as antibodies with higher affinity will bind at lower antigen concentrations.

K_a is also the ratio of the kinetic on- and off-rate constants, which quantify the rates at which a free antibody and free antigen combine (through collisional encounters) to form a binary antibody-antigen complex and at which a binary antibody-antigen complex dissociates to the free antibody and free antigen, respectively.

$$K_\mathrm{a} = \frac{\text{on-rate, } K_\mathrm{on}}{\text{off-rate, } K_\mathrm{off}} \tag{1.3}$$

In addition to affinity, the notion of avidity is crucial for describing the strength of antibody-antigen binding. Avidity is defined as the combined strength of multiple bond interactions. IgM is said to have low affinity but high avidity because it has 10 weak binding sites due to its pentameric structure as opposed to the 2 stronger (higher affinity) binding sites of IgG, IgE and IgD.

Enzyme-linked immunosorbent assay (ELISA)[10] enables the determination of the dissociation constant (K_d) of antigen-antibody equilibria in solution [51].

For measuring affinity, surface plasmon resonance (SPR) is a well-established label-free technique that is frequently used not only to detect affinity of protein-protein, protein-ligand or DNA-DNA interactions, but also for retrieving kinetic information, such as K_on and K_off, on antibody-antigen binding by following the SPR signal in real time [52, 53].

1.4.4 Polyspecificity of antibodies and completeness of the antibody repertoire

Antibodies, just like other proteins, are not monospecific [54]: they are proteins, which bind with varying affinity to a multitude of structures [55–58]. While antibodies are able to bind multiple antigens with comparable high affinity [54, 59–62], this does not necessarily mean the bound antigens are structurally close [47]. The search for the single correct antigen for a given antibody is thus rendered meaningless by the polyspecificity of antibodies [63]—the range of shared specificities is the key observation [64].

The number of B cells (not distinct B-cell clones) present at any one time is estimated to be 10^8–10^9 in mice and 10^{12} in humans [65]. This number is much lower than the number of all possible antigens. Polyspecificity [63] of antibody molecules may thus ensure the completeness of the antibody repertoire [66–69] describing its ability to react to all possible antigens.

[10] ELISA is a technique that essentially requires any ligating reagent that can be immobilized on the solid phase along with a detection reagent that will bind specifically. An enzyme is used to generate a signal that can be quantified. ELISA is also a common means for determining antibody titers. An antibody titer is a measurement of how much antibody an organism has produced that recognizes a particular antigen, expressed as the greatest dilution that still gives a positive result [2].

1.4.5 Humoral specificity and current definitions of specificity

The so called specificity of antibodies is a hallmark of antibody reactivity. It is often described as a selective reaction tailored to a specific cause: clearing of dead cells, pathogens, etc. [70].

A general definition of specificity reflecting its relative nature has been given by Neil S. Greenspan [50]: "Monovalent affinity can be defined as a ΔG (change in free energy) value pertaining to a particular receptor-ligand interaction and specificity can be defined by $\Delta\Delta G$ values that characterize two or more receptor-ligand interactions." Thus, the definition of specificity depends on a frame of reference: the local environment of the studied species has to be given[11] as well as the chemical species to which the studied species is compared to [50].

Greenspan's definition is in its nature consistent with that of van Regenmortel who states that "[...] a perfect fit between epitope and paratope is not a meaningful concept. The degree of specificity of an interaction cannot be linked directly to the size of the antibody affinity constant, and it is generally more meaningful to compare specific interactions in terms of their discrimination potential. The same antibody may thus be called specific or nonspecific, depending on what the investigator is trying to achieve" [12].

Specificity is primarily discussed with respect to monoclonal antibodies [72, 73]. Also, concepts such as cross-reactivity are mostly looked at from a monoclonal antibody's point of view [74–76], whereas, in fact, *monoclonal* specificity does not per se explain *humoral* specificity. Indeed, despite antibody polyspecificity (Section 1.4.4), the population of serum antibodies shows a high degree of specificity towards the eliciting antigen [59]. Serum specificity provides the very basis for the clearing of pathogenic agents from the body. Talmage suggested that "in a mixture of a large number of different globulin molecules, the dominant reactivity will be that common to the largest number of molecules present" [77]. The exquisite specificity of an immune serum could therefore be regarded as an ensemble phenomenon of serum antibodies [59, 71, 78].

1.5 The shaping of the B-cell receptor repertoire

Virtually any substance can be the target of a humoral immune response due to the antibody repertoire's high diversity. The response to even a simple antigen bearing a single antigenic determinant is diverse, comprising many different antibodies, each with a subtly different and unique antigen affinity (Section 1.4.4). The number of different antibodies available at any one time to an individual depends on the number of B cells in an individual as well as to various other factors such as health status, number of antigen encounters in life etc. The diversity of the BCR repertoire is generated by four main processes.

[11] „Solute-solvent interactions, molecular crowding and confinement not directly related to the details of the intermolecular interface can play crucial roles in determining both intrinsic affinity and differential intrinsic affinity." ([71])

1.5.1 Somatic recombination of Ig genes: a mechanism creating threefold diversity

Gene rearrangement takes place during development of B cells in the bone marrow combining two or three gene segments to form a complete V region exon. The gene rearrangement is also referred to as somatic recombination. Three separate loci encode the two Ig light chains (Igκ, Igλ) and the Ig heavy chain. Each light chain locus is composed of three different clusters of gene segments, referred to as variable (V), constant (C) and joining (J) gene segments. The IgH locus bears an additional cluster of diversity (D) gene segments. Somatic recombination thus generates diversity in two ways: first, there are multiple types of copies of each gene segment and different combinations of gene segments can be used in different rearrangement events. This combinatorial diversity is responsible for a substantial portion of the diversity of the heavy and light chain V regions. Second, junctional diversity is introduced at the joints of the different gene segments as a result of addition and subtraction of nucleotides by the recombination process. A third source of diversity originates from combinatorial events, arising from the many different combinations of heavy- and light chain V region pairings forming the antigen-binding site in the Ig molecule. These three mechanisms give rise to a potential diversity of about 10^{12}–10^{13} different BCRs in humans and 10^9 in mice [79]. They take place during the initial development of B cells in the primary lymphoid organs.

The human and mice species generate their BCR diversity in a similar fashion [80]. Prior to the antigenic challenge, these species produce a primary repertoire through the recombination of multiple germline genes [81–83]. However, even though human and mouse antibodies are similar with respect to their diversification strategies, they differ in the extent to which κ and λ light chains are present in their variable light chain repertoires. While the Igκ-V germline genes are dominating the response in mice (95% or more), they comprise only 60% in humans [84].

1.5.2 Somatic hypermutation—a fourth process increasing the diversity of the BCR repertoire

Somatic hypermutation (SHM) introduces point mutations into rearranged V regions of activated B cells, creating further diversity that can be selected for enhanced antigen binding. SHM takes place in the germinal centers (GCs) [85] in the peripheral lymphoid organs after functional Ig genes have been assembled. It introduces point mutations at a rate of $10^{-3}\mathrm{bp}^{-1}\mathrm{generation}^{-1}$ giving rise to mutated BCRs on the surfaces of B cells [81, 86, 87]. In mice and humans, SHM occurs only when B cells respond to antigen along with signals from activated T cells (T-cell dependent B-cell activation) [88]. The Ig C region and other genes are mostly not affected, whereas the rearranged V_H and V_L genes are mutated even if they are non-productive and are not expressed. The base changes are distributed throughout the V region, but are not entirely random due to the existence of certain mutational hotspots [89, 90]. Some of the mutant Ig molecules bind antigen better than the original BCR and B cells expressing them are preferentially selected to mature into antibody-secreting plasma cells or memory B cells. This gives

rise to a phenomenon called affinity maturation[12].

1.5.3 B-cell receptor repertoire analyses

The BCR repertoire is highly variable and of broad chemical diversity and high selectivity (Sections 1.4.2 and 1.5). However, it is still unclear which fraction of the potential repertoire is expressed in an individual at any point in time and how similar repertoires are between individuals who have lived in similar environments [77, 92, 93].

Recently, genome deep-sequencing[13] technologies allowed the exploration of the BCR repertoire due to the recent development of techniques and the exponential reduction in cost of sequencing [94]. The process of obtaining the BCR repertoire starts with the B-cell isolation from the relevant biological sample. Subsequently recombined sequence regions are isolated and sequenced on parallel sequencing machines [95, 96]. According to output sequences, clones are quantified. The use of RNA is among others a source of bias. There are different quantities of mRNA in different cells: active B cells and ASCs produce much higher amounts of mRNA compared with resting B cells.

Weinstein and colleagues sequenced the IgM-BCR repertoire of healthy zebrafish. They discovered that (i) the abundance distributions of both the VDJ repertoire and antibody heavy-chain diversity were similar between individuals, (ii) that VDJ usage is not uniform, (iii) and that individuals can have highly correlated VDJ repertoires [97, 98]. Similar characteristics were also found within the IgM repertoire of human blood cord cells [99]. Furthermore, Weinstein and colleagues used their data to estimate the number of different B-cell clones to be between 1200 and 3500 per fish [93]. For the human system, Glanville and colleagues determined the total diversity of IgM BCRs of peripheral blood mononuclear cells to be at least 3.5×10^{10} per individual [100]. A similar number was reported by Arnaout and colleagues [101].

1.6 Serum antibodies

The serum is a component of the blood, containing neither blood cells nor clotting factors[14]. However, it contains all the electrolytes, antibodies, antigens, hormones, and any exogenous substances (e.g. drugs and microorganisms) [102].

Serum antibodies constitute the antibody repertoire (AR)—the ensemble of secreted antibodies found in the blood at any one time [103, 104]. ASCs can synthesize and secrete several thousand antibody molecules per second [105, 106]. The antibody levels in the serum and other body fluids are maintained by a relatively small population of ASCs making up only about 0.1% to 1.0% of the cells of secondary lymphoid organs and the bone marrow [107–110]. The half-life of antibody molecules in serum is less than 3 weeks [111]. The maintenance of serum antibody levels requires therefore continuous secretion of antibodies [112].

[12] The concept of affinity maturation remains a matter of discussion [70, 88, 91].
[13] Also called *next generation sequencing*, *immunosequencing* or *repertoire sequencing* [94].
[14] Serum is equivalent to plasma after removal of clotting factors.

1.6.1 Antibody isotypes in serum

The serum of human immunocompetent donors mainly contain antibodies of the IgG, IgA, and IgM classes [2]. IgD and IgE are present in serum at only low concentrations, together accounting for less than 1% of total serum Ig. Accounting for about 85% of total serum antibody levels in humans, antibodies of the IgG subclasses are most abundant. IgA abundance amounts to 7%–15% of serum antibodies. Most IgA is secreted as a dimer within mucosal fluids [113]. Roughly 5% of serum antibody is IgM, mainly in pentameric form [112, 114, 115].

IgM and IgG antibodies are already present in serum of newborns before they have contacted any pathogens. IgG levels in fetal serum are comparable to IgG levels of the mother [112]. A fraction of these antibodies is produced by ASCs, which developed from B-1 lymphocytes (Section 1.6.2, [116]).

1.6.2 Antibody secreting cells

Serum antibodies are derived from different types of ASCs, reflecting the dual role of B cells in both innate and adaptive immunity.

The *innate* part is carried out by ASCs of the B-1 lineage which express antibodies that bind often to microbial structures shared by a variety of pathogens throughout the life of the individual [117–119]. Antibodies secreted by B-1 cells are usually of the IgM, IgA, or IgG_3 subclass. B-1-derived ASCs, producing "natural antibodies", are already prenatally active. B-1 cells are present in low numbers in the lymph nodes and spleen (in the marginal zone [120]) and are instead found predominantly in the peritoneal and pleural cavities [112, 121]. B-1 lymphocytes differ from other B lymphocyte subsets in that they arise early in ontogeny, they use a distinctive and limited set of gene rearrangements to make their receptors and they are self-renewing in the periphery. They cannot be boosted: after repeated exposure to the same antigen, they elicit similar, or decreased, responses with each exposure. B-1 B cells, in the mouse, can be further subdivided into B-1a ($CD5^+$) and B-1b ($CD5^-$) subtypes. Unlike B-1a B cells, the B-1b subtype can be generated from precursors in the adult bone marrow [122].

In response to antigen, ASCs also develop from B-2 lymphocytes, in an *adaptive* humoral immune reaction that peaks at about 1 to 2 weeks after antigenic challenge [112, 123]. When naïve B cells traffic through secondary lymphoid tissues and encounter foreign antigen, they can differentiate into multiple fates depending on the type, strength and timing of signals received within the lymphoid microenvironment. Both T-cell-independent (TI) and T-cell-dependent (TD) antigens induce naïve B cells to become short-lived antibody-secreting plasmablasts that localize to extrafollicular regions of lymphoid tissues [124, 125]. TD antigens also induce naïve B cells to seed GCs in lymphoid follicles. Within GCs, B cells undergo SHM, isotype switching and affinity-based selection, which is thought to result in the generation of long-lived memory and PCs [70, 91, 126–130]. Long-lived memory B cells and PCs then migrate from the GC to distinct sites, such as the splenic red pulp, medullary cords of lymph nodes or mucosal-associated lymphoid tissues of the gut for PCs, or splenic marginal zone or

tonsillar epithelium for memory B cells [124, 131]. Alternatively, the cells can egress from their tissue of origin, enter the circulation and take up residence in distal sites (e.g. the bone marrow), so called niches, where they receive survival cues from neighboring cells. During inflammatory or autoimmune responses, PCs can also home to inflamed tissues [132–134].

Of outstanding interest is a study by Bachmann and colleagues who set out to enumerate the number of specific ASCs after immunization with the vesicular stomatitis virus (VSV). VSV is thus an example of a pathogen for which a single epitope is immunodominant. Bachmann and colleagues show that after a certain time a relatively small part of the ASC repertoire is enough to uphold protective antibody titers against VSV: during the acute phase (day 8) of the immune response, more than 50% of all IgG_{2a}-producing ASCs were specific for VSV. In a later phase (days 21 or 50), 10 to 20 times fewer VSV-specific ASCs were present, corresponding to a frequency of approximately $1{:}10^4$ spleen cells [135]. During the memory phase of the anti-VSV response usually 10^4 ASCs/mouse are engaged to maintain a high level of memory IgG against the neutralizing determinant on VSV suggesting a neutralizing anti-viral protective memory-ASC repertoire of 10^2 to 10^4 different VSV-affinities [135].

1.6.3 Antibody repertoire analyses

The recent *BCR repertoire* sequencing approaches (Section 1.5.3) only draw a fragmentary picture of the shape of the *AR*. In fact, it is unknown how the structure of the BCR repertoire [93, 97, 98] compares to that of the AR. The exclusive sequencing of ASCs, which would give a better picture of the AR, is rendered difficult due to their widespread localization throughout the body (Section 1.6.2). A deeper understanding of the AR is of considerable importance as it represents the very foundation of serodiagnostic approaches (Section 1.8) and many B-cell epitope approaches (Section 1.7.3). Brissac and colleagues estimated the number of different IgM specificities present at any one time in murine blood to be of the order to 10^4 [103].

1.7 Studying antibody-peptide binding

1.7.1 Structural and thermodynamic affinity mapping of antibody-antigen interfaces

The most accurate experimental method for the determination of the structure of antigen-antibody complexes is X-ray crystallography [136]. Since the first X-ray crystallographic structure determinations, sequences of immunoglobulin light and heavy chain variable regions have accumulated at an ever increasing rate. Nevertheless, protein crystallography is limited by two main factors: the time required to collect, process, and refine an X-ray data set and the intractability of certain proteins to crystallographic analysis. Currently, the only experimental alternative to X-ray techniques is nuclear magnetic resonance (NMR) which has its own particular shortcomings [137].

Functional analyses of antibody-antigen interaction complement structural analyses in that they describe antibody-antigen interaction by the kinetic rate constants, equilibrium constants (Section 1.4.3), and thermodynamic binding parameters of the complex (Equations 1.4 and 1.5) [26, 138]. The change in enthalpy ΔH and entropy ΔS in combination allow for the calculation of the change in free enthalpy (or Gibbs energy) ΔG, which in turn allows for the determination of the equilibrium association constant K_a.

$$\Delta G = \Delta H - T\Delta S \tag{1.4}$$

$$\Delta G = -RT \log\left(K_\mathrm{a}\right) \tag{1.5}$$

The binding enthalpy primarily reflects the strength of the interactions of the ligand with the target protein (e.g. van der Waals, hydrogen bonds, etc.) relative to those existing with the solvent. The entropy change, on the other hand, mainly reflects two contributions: changes in solvation entropy and changes in conformational entropy. Upon binding, desolvation occurs, water is released and a gain in solvent entropy is observed. This gain is particularly notable for hydrophobic groups [139]. Because the enthalpic and entropic components are related to structural parameters, they can be used (i) as a guide to molecular (drug) design, (ii) as a way to validate structure-based computational predictions of binding energetics (iii) to develop rigorous structure-energy correlations [139]. Calorimetric instrumentation that can be used to directly determine binding enthalpies has become increasingly available over the last two decades [43, 140, 141].

Isothermic titration calorimetry (ITC) is a quantitative technique that can directly measure the binding affinity (K_a), enthalpy changes (ΔH), and binding stoichiometry (n) of the interaction between two or more molecules in solution. From these initial measurements the Gibbs energy (ΔG) and entropy changes (ΔS) can be determined using Equations 1.4 and 1.5 [142]. ITC does not rely on the presence of chromophores or fluorophores (label-free) nor does it require an enzymatic assay [143].

1.7.2 Modeling antibody-peptide binding

Presently, mathematical modeling approaches complement experimental methods of antibody-antigen characterization in at least two interdependent regards: (i) mathematical modeling functions as a generator of hypotheses for experimental studies. (ii) Modeling approaches aim to generalize experimental findings, since a complete characterization of a given antibody-ligand pair would be both money and time consuming.

String modeling

To mathematically simulate binding of antibodies to antigen, bit strings are often used to represent antibodies as well as antigens [68, 144–147]. Antibody and antigen sequences are only composed of two “amino acids”, 0 and 1. The patterns of the bits represent the shapes of molecules and determine their ability to bind with other molecules [144]. In the bit string universe conceived by Farmer and colleagues, molecular binding takes place

when the antibody bit string and the antigen bit string "match" each other. A match occurs when the antigen and antibody have complementary binary patterns [145].

Bit string models are often not only used to study antibody-antigen binding per se. Instead they serve as a means to address broader questions concerning evolution, adaptation and pattern recognition of immune systems [147]. Lancet and colleagues used bit strings to predict sizes of various receptor repertoires [146].

Bit string approaches were also explored in vitro: Fellouse and colleagues [148] obtained functional antibodies from a library of antigen-binding sites generated by a binary code restricted to tyrosine and serine. An antibody raised against human vascular endothelial growth factor recognized the antigen with high affinity and high specificity in cell-based assays.

Molecular modeling

There have been many attempts to design models of antibody combining sites to overcome the experimental limitations (Section 1.7.1). The availability of accurate and reliable modeling frameworks would allow the prediction of the effects of site-directed mutagenesis experiments and enable the intelligent application thereof as well as larger modifications to the combining site (CDR replacement, introduction of catalytic activity, and metal binding sites) and, eventually, tailoring of combining sites to new antigens by means of antibody-antigen docking simulations [137, 149]. Moreover, antibody structures can be used to guide rational efforts to enhance stability [150, 151] or to humanize sequences to minimize immunological response [152, 153].

The Protein Data Bank (PDB) [154] currently contains the structures of around 1000 immunoglobulins and enables the creation of good models for the majority of antibody structures via homology modeling [155–157]. However, due to the high variability of the CDRs, these regions are predicted far less accurately [155]. This is unfortunate, since they are the principal contributors to antibody-antigen binding along with the relative orientation of the antibody light and heavy chains [158]. The accurate prediction of the conformations of CDRs is vital for the understanding of antibody-antigen complexes and has increased in importance with the rise of therapeutic antibodies in healthcare [159, 160]. Despite the high sequence diversity of CDR regions, five of the six CDRs (L1, L2, L3, H1 and H2) are thought to have a set of limited structural conformations (canonical structures) [161]. Reasonably accurate predictions can be made for these five non-CDR-H3 regions using a set of sequence based canonical rules [162, 163]. More recently, the canonical structures have been updated and it was shown that non-CDR-H3 regions are largely predictable ($\approx 85\%$) using sequence, gene source and framework regions [164, 165].

Recently, online web servers for high-resolution molecular homology modeling, based on antibody sequence data, such as Web Antibody Modeling [166], Prediction of Immunoglobulin Structure [167], as well as Rosetta Antibody became available: not only are these web-based solutions directed at yielding highly accurate predictions of antibody binding sites, but also to rendering in silico antibody-antigen docking experiments[15]

[15]Docking is a computational technique that samples conformations of small molecules in protein binding

possible [155].

Quantitative structure modeling

Quantitative structure-activity relationship (QSAR) models are regression models used in the chemical and biological sciences and engineering. QSAR models relate measurements of a set of "predictor" variables to the behavior of the "response" variable. The predictor variables consist of the properties of chemicals whereas the response variables represent the chemicals' activity. The analysis of QSARs is performed to model ligand-binding site interactions, thus being of use in the analysis of molecular recognition [12].

Proteochemometrics is derived from chemometrics[16] and is related to QSAR. In proteochemometrics, simultaneous modifications of all interacting molecules are applied, whereas in QSAR only the interacting ligand is modified. In proteochemometrics, descriptions of both proteins and the interacting ligands are correlated to experimentally measured interaction data by applying multivariate data analysis [73]. Mandrika and colleagues used a proteochemometric approach to analyze the amino acids and amino acid physico-chemical properties that are involved in antibody recognition of peptide antigens. To this end, they used a study system comprising a diverse single chain antibody library derived from the murine monoclonal antibody anti-p24 (HIV-1) CB4-1. The library was manufactured by SPOT synthesis [40]. The binding pattern obtained was correlated to physico-chemical descriptors (z-scales) of antibody and peptide amino acids using partial least-squares projections to latent structures (Section 3.7). The authors claim that with this approach, the physico-chemical properties of each interacting amino acid residue of both the peptides and the antibodies being essential for the antigen-antibody recognition could be retrieved from the model [73].

1.7.3 Predicting antibody-peptide binding

The ability to predict B-cell epitopes for a given protein is a precursor to new vaccine design and diagnostics [30][17] (Section 1.7.1).

Epitope prediction is often done by immunizing the host with the antigen in question followed by profiling the resulting serum-antibody response [172–175] with a wide array of methods some of which are shortly outlined in this section.

Although it is believed that the majority of B-cell epitopes are discontinuous epitopes [34] (Section 1.4.1), the experimental determination of epitopes has focused primarily on

sites. It uses scoring functions to assess which of these conformations best complement the protein binding site [168]. Warren and colleagues conclude in their survey of docking programs that "all of the docking programs were able to generate ligand conformations similar to crystallographically determined protein/ligand complex structures for at least one of the targets". However, "no single program performed well for all of the targets. For prediction of compound affinity, none of the docking programs or scoring functions made a useful prediction of ligand binding affinity" [168].

[16] Chemometrics is the science of extracting information from chemical systems by data-driven means [169].

[17] Recently, the development of peptide-based vaccines, relying on advances in the prediction of linear epitopes, has been suggested to be severely limited by a narrow reductionist view of vaccine design, ignoring that the majority of epitopes is discontinuous [30, 34, 170–172].

the identification of continuous B-cell epitopes [176, 177].

Apart from X-ray crystallography, which represents a structural approach to epitope mapping [33, 45, 178, 179] (Section 1.7.1), one important experimental technique to map epitopes is the use of phage-display libraries [180]. By selecting phages from a library for their ability to bind antibodies specific for a known antigen, linear peptide sequences that cross-react with these antibodies, commonly referred to as mimotopes [181], can be discovered [182].

Another approach to epitope mapping of continuous epitopes is the screening libraries of short synthetic peptides that span the entire target antigen [173, 183–186]. Overlapping peptides for mapping sequential epitopes cannot only be synthesized on pins, but also on a cellulose membrane support [187] and microarrays [188–190] (Section 1.8.1). Other technologies for B-cell epitope mapping include fragmentation methods, competition methods and antigen modification methods (site-directed mutagenesis) [178].

In addition to molecular modeling (Section 1.7.2) and experimental epitope mapping approaches, several computational methods for prediction of continuous B-cell epitopes have been published in recent years.

Propensity scale methods assign a propensity value (scores) to each amino acid which measures the tendency of an amino acid to be part of a B-cell epitope (as compared to the background)—an approach similar to the used z-scales in proteochemometric modeling [191] (Section 1.7.2). These methods rely on the observed correlations between specific physico-chemical properties of amino acids and the antigenic determinants in protein sequences to identify the location of the linear B-cell epitopes in the query protein sequence. The propensity scores are used as a basis for predicting whether a given amino acid sequence residue is likely to be part of a linear B-cell epitope. The first propensity scale method for predicting linear B-cell epitopes was introduced by Hopp and Woods [192] and utilized the Levitt hydrophilicity scale [193]. The Levitt scale is based on the assumption that antigenic determinants of protein sequences correspond typically to sequence windows that contain a large number of charged and polar residues and lack large hydrophilic residues[18]. Subsequently, several other propensity scales based on hydrophilicity [194], flexibility [195], turns [196] and accessibility [197] have been proposed for predicting linear B-cell epitopes. Approaches combining different scales were also published [198–200].

Recently, Blythe and Flower [172] have performed an assessment of 484 amino acid propensity scales to examine the correlation between propensity scale-based profiles and the location of linear B-cell epitopes in a data set of 50 proteins. Their study found that even the best combinations of amino acid propensities yielded B-cell epitope predictions that were only marginally better than random [177].

Due to the poor results yielded by propensity scales alone, several authors have explored methods for improving the predictive performance of propensity scale methods by combining them with machine learning methods such as Hidden Markov models or support vector machines [201, 202]. However, the combination of scales with several

[18] A study published by Lollier and colleagues [174] suggests that hydrophilicity of amino acid residues may not play a primary role per se in epitope definition.

machine learning algorithms showed little improvement over single scale-based methods [30, 182].

The increasing number of experimentally characterized linear amino acid sequences of B-cell epitopes [203] led several authors to explore approaches using exclusively machine learning based methods for predicting linear B-cell epitopes [204–206]. All these approaches showed prediction accuracies between 70% and 80%. However, benchmark studies show that the predictive performance of machine learning methods is also rather poor [30].

Prediction of conformational epitopes has gained traction over the recent years due to an increasing amount of sequence and structure data which is being stored in various databases making it possible to use machine learning approaches for prediction. A detailed overview over recent developments is given in a review from EL-Manzalawy and Honavar [177].

In conclusion, even though publications involving epitope prediction approaches are multiplying owing to steadily increasing computational power and databases, the predictive performance of current methods is far from ideal [172, 182]. Furthermore, immunogenicity of proteins is poorly understood [207], and it remains an open question whether B-cell epitopes could after all be deciphered as intrinsic features of proteins [30, 177].

1.8 Serological diagnostics with antibody profiling

The discovery of humoral antitoxic antibodies in the early 1890s exerted a profound influence upon the future development of both immunologic practice and thought [208]. The demonstration of the presence of specific agents in the serum of immunized animals opened the way for the development of such serologic tests as agglutination, the precipitin reaction, and complement fixation [208]. More recently, techniques such as ELISA joined the ranks of serological methods serving to diagnose diseases such as HIV and Hepatitis C [208]. Specific antibody-antigen recognition is therefore the basis for the widespread use of antibodies for molecular identification in research and in the clinic [38, 208].

As the AR (Section 1.2) is subject to constant change—both with respect to respective antibody concentrations and overall antibody composition [135]—due to continuous antigen encounter and the establishment of immunological memory [3], its investigation provides the possibility to gather information about both past and on-going immune responses, and ultimately about the immune state of the body [209]. Antibodies are rather easily detectable and amplify the immune response [210].

Due to the AR's high diversity, high-throughput immunoblot [211, 212], SEREX [213], phage display [180, 214–216] and microarray technologies using as probe molecules among others glycans [217, 218], aptamers [219–221], peptides (Figure 1.2) and proteins (Section 1.8.1) have been used for large-scale profiling of serum antibody binding.

Such antibody profiling data were used for disease classification by exploiting differences of antibody-peptide binding patterns (Table 1.1) as well as for characterizing antibody binding patterns in general [222]. Ultimately all these approaches harness the antibodyome's [223] specificity potential to diagnose diseases in a minimally invasive way

based on the premise that the AR reflects an individual's health status [210].

Thus, the underlying data-analytic assumption of serodiagnostic approaches is that disease-induced antibodies dominate antibody-ligand binding (ligand: glycans, aptamers, proteins, peptides etc.) in a *consistent* fashion across infected individuals: only consistent antibody binding patterns allow the discrimination of healthy and diseased individuals.

1.8.1 Antibody profiling with peptide microarrays

The use of high-throughput-screening methods made it obvious that antibody binding *patterns*, that is, signal intensities (Section 1.8.1) of multiple peptides, are rather the more important observation of antibody profiling studies and not necessarily single signals. Patterns show a higher discriminatory power regarding the separation of healthy and diseased groups [188, 224]. In this regard, especially random-sequence peptide array approaches have emerged as a tool of choice for antibody profiling. In the following, I will present the production and functioning of peptide[19] microarrays, one of the major platforms for antibody profiling [223].

Production of peptide microarrays

A microarray is a gridded presentation of molecules across a planar surface. Each molecule occupies a pre-defined position on the grid, denoted as a spot, which encodes its identity [225]. Microarray surface types include chemically derivatized planar glass or silicon chips, flow cytometric microbead assays, arrays in plastic microwells, nitrocellulose on chips and three-dimensional gels [226].

The production of peptide microarrays involves the covalent linkage of peptides to the microarray surface. To enable this linkage, slides are chemically treated [227] with compounds such as aldehyde-containing silane reagents. The aldehydes react readily with primary amines on the proteins to form a Schiff's base linkage [228]. To fabricate peptide microarrays, printing robots deliver nanoliter volumes of peptide samples to the slides. Printing approaches include contact- as well as non-contact printing [228, 229]. The disadvantages of glass or silicon chips lie in the batch-to-batch variability associated with attachment chemistry [226].

Signal detection of antibody-peptide binding

For detecting antibody-peptide binding, label-free approaches such as SPR (Section 1.4.3) and labeled-probe approaches exist. The latter include labeling by a chromogen, a fluorophore or a radioactive isotope. Fluorescence detection can be by direct labeling of monoclonal antibodies which then bind to the spotted peptides or by indirect approaches via secondary antibody detection, where an isotype-specific antibody is fluorescently labeled [230] to bind to, for instance, serum antibodies (Figure 1.2). The resulting fluorescence signal intensity is measured with a scanning device (e.g. laser scanner, Figure 1.2). The fluorescence is proportional to the amount of bound antigen [231, 232].

[19] For improved readability, I will use the term peptide for both proteins and peptides.

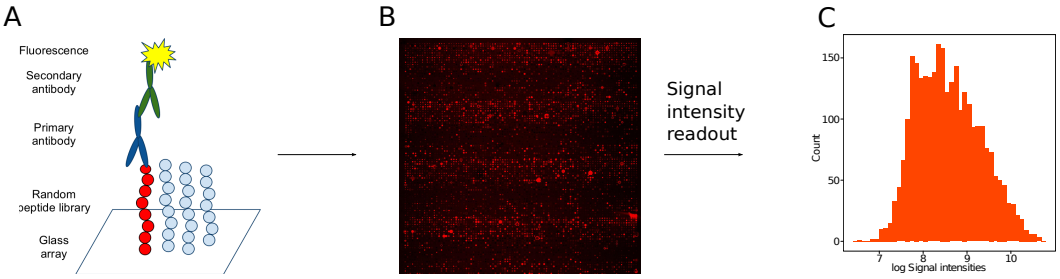


Figure 1.2: Overview over the workflow of antibody profiling with random-sequence peptide arrays. (A) Incubation of a random-sequence peptide array with primary antibody (monoclonal or serum antibodies), of which the binding is detected with a fluorescently-labeled secondary antibody. (B) The incubated peptide array is scanned yielding an image in which the spot intensities (here in red) are proportional to antibody binding to the peptide spots. (C) The signal intensity read-out (also termed "antibody binding profile" or "signal intensity profile") yields a signal intensity distribution (here log-transformed), which is used for downstream analyses such as serological diagnostics (Section 1.8) or B-cell epitope mapping (Section 1.7.3). The presented workflow is adopted by the majority of antibody profiling studies (Table 1.1).

State of the art of antibody profiling with peptide microarrays

Antibody profiling studies with peptide microarray differ widely with respect to methodology and objectives. Among those shown in Table 1.1, two types of approaches can be discerned.

Type 1 Peptide microarray approaches which, presuming humoral specificity, interpret measured signal intensities in *absolute* terms: in addition to binding patterns, much attention is given to absolute measured peptide signal intensities, which are interpreted as a reflection of the peptides' function (eliciting antigen) in the studied disease. These approaches, therefore, mostly use disease-dedicated peptide libraries to find potential disease-antigens[20]. Articles of the survey (Table 1.1) belonging to this category are Hueber et al. [234], Robinson et al. [235] and [236], Gaseitsiwe et al. [237] and Merbl et al. [224].

Type 2 Peptide microarray approaches which, presuming humoral specificity, interpret measured signal intensities in *relative* terms: peptide signal intensities are analyzed as consequences of antibody polyspecificity [63] (Section 1.4). The relative differences in binding patterns between the healthy and the diseased case represent the main finding. The fact that polyspecificity of antibodies renders the nature of the eliciting antigen(s) unimportant makes random-sequence peptide arrays, which are primarily used in this type of approach, a relatively cheap, unbiased, non-pathogen-restricted, and user-friendly tool for serological diagnostics [189, 238]. Articles of the survey (Table 1.1) belonging to this category are Legutki et al. [189], Reddy et al. [239] and Bongartz et al. [188].

[20] This assumption is problematic due to antibody polyspecificity (Section 1.4.4). See also Kroening and colleagues for further evidence against this hypothesis [233].

Table 1.1: Literature survey of studies using antibody profiling for serological diagnostics. Legend of abbreviations, which are not listed in the *List of Abbreviations*. AAA: autoantigen microarray, AA: autoantigen, AD: Alzheimer disease, BCResp, B-cell response, EAE: Experimental autoimmune encephalomyelitis, FC: Fold change, FL: Fluorescence, FP: False positive, H: Human, HB: *Heligmosomoides bakeri*, HC: Healthy control, IR: Immune response, LDA: Linear discriminant analysis, M: Mouse, NA: not available, NRP: non-random peptides, P: proteins, PA: peptide array, RM: regression model, R: Rat, RA: Rheumatoid arthritis, RPA: random peptide array, SAM/PAM: Significance (Prediction) analysis of microarrays, SI profile: signal intensity profile, SP: Spots, SVM: Support vector machine, TB: Tuberculosis.

Study			**Array platform**		**Sample processing**		**Data processing**		
Ref.	Objective	Results	Number of peptides	Ig-Isotype-detection	Sample type	Sample dilution	Signal definition	Normalization	Bioinformatical methods used
[234]	Ident. of distinct serum SI profiles in patients with RA.	Classification of different RA patient groups with AA is possible.	225 NRP and Pr from the synovial proteome.	IgM, IgG	H (27 RA patients, 11 HC).	1:150	Median of the FL of the replicated SP.	Scaling to a positive control (IgM).	SAM and PAM.
[235]	Measurement of serum SI profiles with structurally diverse AA.	Description of AAA technology.	196 distinct putative AA on 1152-feature arrays.	IgM/IgG	H (50 samples of SLE patients).	1:150	Median of the background-subtracted FL of replicated SP.	Scaling to a positive control (IgG).	Simple SI analysis.
[236]	Measurement of EAE-autoAb responses.	AAA-identified EAE targets functioned to treating established EAE. Reduction of epitope spreading of autoreative BCResp.	2304-feature myelin proteome arrays containing $\leq$ 232 distinct antigens.	IgM, IgG	M/R (Number: NA).	1:150	Median of the background-subtracted FL of replicated SP.	Scaling to a positive control (IgG).	SAM.
[189]	Profiling of the IR with RPA.	Sets of peptides discriminated healthy and influenza-infected M groups as well as H individuals.	10000 20-mers	IgM, IgG_1, IgG_{2a}, IgG_3	M (Serum samples, number: NA), (H, Serum samples, number: NA).	1:500	Average of triplicates of median SI per SP.	Each array was normalized to the 50th percentile. SIs of less than 0.01 were set to 0.01.	PCA.

Continued on next page

Table 1.1 – *Continued from previous page*

Ref.	Objective	Results	Number of peptides	Ig-Isotype-detection	Sample type	Sample dilution	Signal definition	Normalization	Bioinformatical methods used
[239]	Measurement of HC and case serum SI profiles with unnatural molecules.	Identification of two candidate IgG biomarkers for AD.	Two copies of 4608 octameric (AA are non-naturally occuring) peptoids and control SP.	IgG	H (Serum samples (22 HC, 22AD), M, Serum samples: HC: NA, EAE-induced (via MOG injection) C57BL/6 M: 30	15 µg/ml	Local background subtracted median SP SI ($40000 > SI > 10000$) were used for further analysis.	None	FC analysis.
[188]	Discrimination of RPA serum SI profiles with few peptides.	Classification of different M strains and HB-infection with high accuracy and small peptide sets.	249 different random-sequence 14-mers.	IgM	M, Serum samples: 15 HC (C57BL/6), 28 (HC, case, BALB/c).	1:10	Elimination of FP SIs.	None	PCA, LDA, P-SVM.
[237]	Serum Ab-based target identification of TB antigens using high-content PA.	Successful classification of HC and TB patients.	7776 peptide SP.	IgG, IgA	H (Serum samples: 35 HC, 34 TB).	1:100	Elimination of FP SIs.	Use of linear RM to remove effects of slide, sub-array, block and FP.	SAM, PAM.
[224]	Measurement of serum SI profiles of healthy and metastatic inbred C57BL/6 M.	Classification of healthy and cancerous M.	327 AA.	IgM, IgG	M (Serum samples: HCs and 2 groups of 20 M injected with different lung cancer lines).	1:500	Mean log intensity measures of at least 4 replicate SP.	Division of each array by its median SI.	Wilcoxon Rank-sum test, SVM.

Standardization of experimental and bioinformatical approaches to antibody profiling with peptide microarrays

As observed in the literature survey (Table 1.1), both experimental handling (dilution of sera, microarray platform, chosen peptides, etc.) and the data preprocessing is rather non-standardized across studies, which is in contrast to cDNA microarray studies [240–244].

The more recent emergence of peptide arrays as an experimental method for antibody profiling as well as the different nature of data that represent antibody-peptide binding profiles may be the cause for experimental and bioinformatical heterogeneity across studies.

Bioinformatic methods from gene array analysis cannot be readily transferred to the analysis of antibody-peptide binding, because in nucleic acid microarray technologies, binding is essentially only between two types of molecules of complementary sequence. With peptide microarrays used for antibody profiling, signal intensity profiles are continuously distributed and binding is not restricted to a single complementary molecule[21]. Due to antibody polyspecificity (Section 1.4.4), multiple antibodies can bind to the same peptide on the array and a single antibody may also bind to multiple peptides on the array: an issue almost non-existent in gene expression arrays [238, 245, 246].

As suggested by the non-exhaustive survey of peptide array studies (Table 1.1), approaches to antibody profiling are largely phenomenological. However, standardization of experimental and bioinformatical protocols can only be done faithfully if the genesis of antibody binding profiles is understood [238]. Publications focusing on the theoretical study of antibody profiles are, as far as I know, almost non-existent [245]. Consequently, publications [247–251] focusing on the standardization of peptide array approaches are of uncertain theoretical foundation.

Therefore, this thesis sets out to study antibody-peptide binding with a mathematical model. The advantage of this theoretical approach is that the effect on antibody binding profiles of experimentally mostly inaccessible parameters such as antibody composition and diversity can be readily studied.

1.8.2 Characterization of the murine parasite *Heligmosomoides bakeri*

In the following, I will characterize the murine parasite *Heligmosomoides bakeri* (HB), since antibody-peptide reactivity data involving samples of mice attained by an HB-infection are presented in the Results section (Chapters 5–8).

HB typically causes long-lasting infections in mice and in this respect has been a laboratory model of chronic intestinal nematode infections [252–254]. HB is a natural enteric nematode parasite of murine rodents that enters the gastrointestinal tract at larval stage L3 then penetrates the epithelial cell barrier of the small intestine to mature within the submucosa to an L4 stage. Approximately 8–10 days after infection, the parasite

[21] It has recently been shown that antibody-binding to random-sequence peptide arrays is largely driven by the variable region of the antibody: competition with 10-fold excess Fc protein showed no effect on the measured antibody binding profiles [210].

exits the intestinal mucosa to populate the intestinal lumen and establishes a chronic infection as a sexually mature adult producing viable eggs that are secreted through the feces [254–256].

While target antigens for some helminth infects are known [257]—immunogenic antigens of HB are largely unknown—in particular it is not known whether the antigens that induce protective antibodies are derived from the same or different stages of the HB life cycle [258].

Works by Wojciechowski [258], McCoy [256] and colleagues showed that both polyclonal and affinity-matured IgG antibodies play an essential role in protective immunity to HB expulsion (IgM, IgE, and IgA do not play a significant role in resistance). Polyclonal IgG antibodies, present in naïve mice and produced following HB-infection, functioned to limit egg production by adult parasites. Comparatively, affinity-matured parasite-specific IgG antibodies that developed only after multiple infections were required to prevent adult worm development.

2 Objectives

Antibody profiling with random-sequence peptide arrays holds great promises for serological diagnostics [188, 189, 245] as it represents a versatile method for the discrimination of either healthy from diseased individuals or individuals at different disease stages. Serological diagnostics based on antibody profiling rest primarily on the assumption that antibody profiles of diseased and healthy individuals differ consistently from one another. The challenge for antibody profiling is therefore twofold: reflecting the change in the antibody mixture induced by the disease while taking into account the variability of antibody profiles of healthy individuals. To my knowledge, the antibody repertoire's impact on antibody binding profiles has not been extensively investigated. Since the characterizing components of antibody repertoires, such as composition and concentration, are difficult to study in vitro, an aim of this thesis is to provide a mathematical model for antibody-peptide binding.
The study of both the genesis and specificity of antibody binding profiles with a mathematical model and the subsequent validation of the model's predictions by in vitro antibody-peptide reactivity data is the major objective of this thesis.

This work has the following specific aims:

Description of a mathematical model for antibody-peptide binding. This model is based on the law of mass action incorporating as parameters (i) antibody and peptide sequences and (ii) antibody concentrations.

Mathematical analysis of the proposed model and implementation of a framework for simulating antibody-peptide reactivity data. The mathematical analysis will center on studying the impact of both antibody composition and diversity on signal intensity. Furthermore, the simulation framework allows for the study of the impact of parameters such as peptide length and peptide library size, which are somewhat elusive to analytical analysis. It also provides the opportunity to investigate external parameters such as noise.

Test of the predictions generated by both the mathematical model and the simulation framework with in vitro antibody-peptide reactivity data. In order to quantify the generality of the predictions, the studied data sets include various peptide libraries, both human and murine serum samples and monoclonal antibodies.

Discussion of the results obtained with the mathematical model, the simulation framework and the in vitro data in light of their significance for serological diagnostics and B-cell epitope mapping.

3 Methods

This *Methods* chapter is divided into two parts. The first part constitutes the description of the experimental methods (Sections 3.1–3.5) none of which were performed by me. Their knowledge is nevertheless important for the understanding of this work's results. For a more extensive description of experimental methods, please refer to Lück (PhD thesis, [259]). The second part describes the computational methods used in this thesis (Sections 3.6–3.9).

3.1 Peptide microarrays used for incubation with serum or plasma samples

In this work, different peptide microarrays with varying peptide libraries, produced by both contact and non-contact printing technique have been used for the analysis of antibody-peptide binding. Detailed information on peptide libraries and their amino acid composition is provided in the Tables 3.1 and 3.2.

(i) Each peptide spot on a given microarray is assumed to have an equal density of functional groups. (ii) The amino acid sequence of all analyzed peptides is known.

3.1.1 JPT microarrays

Non-contact printed random-sequence peptide microarrays

Non-contact printed peptide microarray slided obtained from JPT Peptide Technologies GmbH (Berlin, Germany) covered five identical sub-arrays each including 255 random-sequence[1] 14-mers and 45 EAE (experimental autoimmune encephalomyelitis) antigen peptides (Figure 3.1). In this work, only the antibody binding to the 255 random peptides was analyzed. Their sequence was designed with a random generator, which did not allow three or more consecutive repeats of an amino acid. This library will hereafter be referred to as $J^{255}_{14\text{-mer}}$.

Each sub-array comprises TAMRA-derived peptides as internal fluorescence control, and mouse-IgM, mouse-IgG, human-IgM, human-IgG, human IgE as secondary antibody controls.

Contact printed random-sequence peptide microarrays

JPT contact-printed microarrays were made of three identical sub-arrays in two different formats. Format 1 constitutes sub-arrays of 7056 spots, which include the analyzed

[1] Henceforth, "random" will stand for "random-sequence".

15-mer random peptide library of 3352 peptides and a non-analyzed non-random 15-mer sub-library (STY-library) containing S, T or Y as the 8th amino acid of each peptide. In addition, human IgM, IgG, IgA and IgE (according to the batch also mouse IgM and IgG) as secondary antibody controls and empty spots were found on arrays of format 1. Analyzed peptide libraries originating from this fabrication will be called $J^{3352}_{\text{15-mer}}$.

Format 2 includes sub-arrays on which 13-mers and 15-mers are displayed. These arrays counted 9600 spots comprising the analyzed 15-mer random peptide library of either 3418 ($J^{3418}_{\text{15-mer}}$) or 3626 ($J^{3626}_{\text{15-mer}}$) peptides[2], an analyzed 13-mer [3] library $J^{2304}_{\text{13-mer}}$, a non-analyzed STY-library containing S, T or Y as the 8th amino acid of a peptide. In addition, human IgM and IgG as secondary antibody controls[4] and empty spots were found on arrays of format 2. Cysteine was not constitutive of any of the peptides in either of the two formats (Table 3.2).

Peptides, produced by spot synthesis[5] [40], are linked to the arrays in an N-terminal fashion. The spot diameter is about 100 µm with an approximate density of functional groups of 15 fmol/mm^2.

3.1.2 Pepscan microarrays

Peptide microarrays obtained from Pepscan (Zuidersluisweg, 28243 RC Lelystad, The Netherlands) are made of three identical sub-arrays each of which shows a total of 1024 spots. Spots include the analyzed library of random 15-mers and shorter peptides, empty spots and human IgM and IgG secondary antibody controls. Cysteine is not constitutive of any of the peptides (Table 3.2). The first four lines of spots of each sub-array were removed prior to the analysis of antibody-peptide reactivity leading to an analyzed number of 942 15-mers ($P^{942}_{\text{15-mer}}$).

According to the data sheet provided by Pepscan, peptide microarrays were spotted onto microarrays with a proprietary surface chemistry, based on a co-polymer of acrylic acid and polyethylene glycol moieties. This surface features a thin (50–100 nm) hydrophilic environment. The peptides were generated by solid-phase synthesis and covalently coupled with a low-charge coupling chemistry.

The microarrays were spotted using a split-pin microarray spotter in a controlled environment. The spot distance (center-to-center) is 560 µm with an approximate density of functional groups 50 fmol/µm^2.

3.2 Incubation of peptide microarrays

Peptide arrays were either manually incubated (Section 3.2.1) or in an automated fashion (Section 3.2.2).

[2] The differing number of analyzed random 15-mers is due to different batches (Table 3.1).
[3] The 13-mer sequences are not random since the 7th position is as a rule S, T or Y.
[4] The secondary antibody controls may or may not be present depending on the batch.
[5] For both JPT (contact and non-contact printed) and Pepscan arrays, details related to array manufacturing are scarce.

Batch number	Analyzed random peptide library	Experimental study
879, 901	$J^{255}_{14\text{-mer}}$	Mouse study (MS)
840	$J^{255}_{14\text{-mer}}$	Monoclonal antibodies
1189	$J^{2304}_{13\text{-mer}}$, $J^{3418}_{15\text{-mer}}$	Glioma 08 study
1233	$J^{3352}_{15\text{-mer}}$	Glioma 09 study
1190	$J^{2304}_{13\text{-mer}}$, $J^{3418}_{15\text{-mer}}$	Slovenian healthy study (SHS)
1133	$J^{2304}_{13\text{-mer}}$, $J^{3418}_{15\text{-mer}}$	NephroFIT study
1027	$J^{2304}_{13\text{-mer}}$, $J^{3626}_{15\text{-mer}}$	NOD study (NS)
0151	$P^{942}_{15\text{-mer}}$	NephroFIT-Pepscan study
0153	$P^{942}_{15\text{-mer}}$	NephrOT study

Table 3.1: Analyzed random peptide libraries are shown with both the associated microarray batch number and experimental study. As to the Mouse study (Section 3.1.2), the repeats used to characterize technological variability were incubated on batch 901, whereas incubations analyzing biological variability were performed on batch 879 (Section 3.5.8). Batch production is a technique used in manufacturing, in which the object in question is created stage by stage over a series of workstations (`http://en.wikipedia.org/wiki/Batch_production`).

After either type of incubation, antibody-peptide binding signals were recorded with a microarray scanner (GenePix 4000B or GenePix 4200AL, Molecular Devices GmbH, Ismaning, Germany) at 532 or 635 nm (10 µm resolution, 5% laser power, 400 photomultiplier excitation) unless mentioned otherwise. Microarray images were stored in a 16-bit TIFF-format.

3.2.1 Manual incubation

The microarrays were briefly immersed in 100% v/v ethanol, washed three times with T-PBS (phosphate buffered saline containing 0.05% w/v Tween20), three times with deionized water and dried by centrifugation. Since the microarray surfaces had been pre-treated to minimize unspecific binding of the target antibodies, no blocking step was required prior to incubation. All incubations were performed using a five-well adhesive incubation chamber (Multiwell GeneFrameTM, ABgene Germany, Hamburg, Germany) with a total assay volume of 45 µl per well. Serum was diluted 1:10 in T-PBS. After incubation for 4 h at room temperature, the microarrays were washed three times with T-PBS and three times with deionized water. Secondary antibodies were diluted in T-PBS (20 µg/ml, 300 µl) and incubated for 1 h at room temperature. The microarrays were washed three times with T-PBS, three times with deionized water, rinsed with running deionized water and dried by centrifugation. Water, ethanol and PBS were filtered.

3.2.2 Automated incubation

Automated incubations were performed using the Tecan HS 4800 hybridization station (Tecan, Mannedorf, Switzerland). Microarrays were washed with T-PBS (washing buffer)

Amino acids	Frequencies of amino acids in analyzed peptide libraries [%]					
	$J^{2304}_{13\text{-mer}}$	$J^{255}_{14\text{-mer}}$	$P^{942}_{15\text{-mer}}$	$J^{3352}_{15\text{-mer}}$	$J^{3418}_{15\text{-mer}}$	$J^{3626}_{15\text{-mer}}$
A	5.0	5.4	5.3	5.2	5.1	5.1
C	0	1.5	0	0	0	0
D	4.6	7.1	5.3	5.3	5.3	5.3
E	4.8	6.0	5.6	5.3	5.3	5.2
F	4.8	4.4	5.3	5.4	5.3	5.2
G	5.3	5.2	5.4	5.3	5.3	5.2
H	4.9	4.0	5.1	5.2	5.2	5.3
I	5.0	3.8	5.2	5.4	5.4	5.5
K	5.1	7.3	5.2	5.4	5.4	5.2
L	5.0	5.2	5.2	5.3	5.2	5.6
M	5.0	2.4	5.3	5.4	5.4	5.2
N	4.8	5.6	5.2	5.3	5.2	5.1
P	5.0	5.1	5.3	5.2	5.2	5.1
Q	4.7	5.0	5.3	5.2	5.2	5.1
R	4.8	6.7	5.1	5.2	5.2	5.3
S	7.2	6.5	5.1	5.1	5.2	5.4
T	9.8	6.2	5.2	5.3	5.3	5.2
V	4.5	5.0	5.1	5.2	5.2	5.1
W	4.7	3.1	5.3	5.2	5.2	5.0
Y	4.8	4.6	5.5	5.2	5.3	5.9

Table 3.2: Frequencies of peptide amino acids by analyzed peptide library. Frequencies may not add up to 100% due to rounding effects. J/P^x_y denotes the array manufacturer (JPT [J], Pepscan [P]), the analyzed number of peptides x and their length y. Except for $J^{255}_{14\text{-mer}}$, cysteine (C) is absent from all analyzed peptide libraries. For further information on the itemized peptide libraries, please refer to Section 3.1.

and blocking buffer (PBS, 1% BSA) and blocked for 30 min at 30°C. After three further washing steps with T-PBS, samples were injected, followed by an array incubation for 2–3 h at 30°C with constant agitation. Thereafter, microarrays were washed four times with T-PBS followed by the injection of the secondary antibody, which was incubated for 1 h at 30°C (contant agitation). Finally, microarrays were washed three times with T-PBS, twice with 0.01x SSC buffer and were rinsed with running deionized water for 3 min.

3.3 Signal detection and determination of raw signal intensities

3.3.1 Signal detection

Signal intensities were quantified with either the GeneSpotter[6] (MicroDiscovery GmbH, Berlin, Germany) or GenePix Pro 6.0 (Molecular Devices GmbH, Ismaning, Germany) software by taking the median pixel intensity of a circular region around the center of each spot [259].

3.3.2 Determination of raw signal intensities

The median signal intensities of each sub-array were averaged to yield one signal intensity value per peptide per array. This averaging procedure was not performed for the $J^{255}_{\text{14-mer}}$ library, since on every sub-array a different mouse serum was applied.

3.4 Preprocessing of in vitro antibody-peptide reactivity data

3.4.1 Preprocessing of antibody-peptide reactivity data prior to signal intensity profile analysis

Fluorescence signal intensities were not normalized unless stated otherwise. If normalized, signal intensities were transformed as detailed in Section 3.4.2.

3.4.2 Preprocessing of antibody-peptide reactivity data prior to AAWS analysis

Measured raw signal intensities were log-transformed ($\log(I)$). Subsequently, the signal arising from the polyclonal secondary antibody was removed according to the linear model:

$$\log(I) = \beta_0 + \beta_1 \log(I_{\text{Secondary Antibody}}) + \epsilon. \tag{3.1}$$

By partial least squares regression (PLSR)-based computation (Section 3.7) of the intercepts, β_0 and β_1, $\log(I)$ was replaced with the resulting PLSR-computed, mean-centered and scaled-to-unit variance residuals ϵ prior to determination of AAWS. The signal of the secondary antibody ($I_{\text{Secondary Antibody}}$) was obtained by incubating an array instead of serum/plasma with blocking buffer (PBS, 1% BSA, "blank", Section 3.5).

3.5 Experimental studies

In the following, experimental methods are grouped by experimental study. The peptide library type and batch number for a given experimental study can be found in Section 3.1 and Tables 3.1 and 3.2. Signal detection was performed, if not stated otherwise, as described in Section 3.3.

[6] Only the antibody-peptide reactivity data of the mouse study was analyzed with GeneSpotter. All other studies were performed with GenePix Pro 6.0.

Technological variability of signal intensity profiles was assessed for each experimental study with repeated measurements (repeats) of healthy control sera. The correlation of repeats was found to be generally above $r_{\text{Pearson}} = 0.85$. However, technological variability *across* studies is high: the Pearson correlation coefficients of blank[7] signal intensity profiles *between* studies was found to be mostly in the range of $0.20 < r < 0.40$ (both for JPT and Pepscan, data not shown): the variability across batches is high [259].

The dependence of the secondary detection antibody on *serum* signal intensity profiles was found to be low: the correlation between blank and serum signal intensity profiles was generally in the range of $0.10 < r_{\text{Pearson}} < 0.30$. In contrast, the impact of the secondary-detection antibody on profiles of *monoclonal* antibodies was found to be large (Figure S.4).

Plasma as well as serum samples were used to measure antibody binding profiles. Reports indicate that "fresh serum, fresh plasma, frozen serum and frozen plasma from the same volunteer showed almost no discernible differences" in antibody binding profiles [210].

In each section, the person who performed the experiment is named[8].

3.5.1 Slovenian healthy study (SHS)

The Slovenian healthy study (SHS) is a baseline study in which 16 healthy human individuals have donated blood at three different time points. The results of this baseline study (peptide array results excluded) were recently published [260].

Incubations, plasma IgM quantification and signal detection were performed by Bodo Steckel.

Slovenian healthy study: Plasma samples

Peripheral blood samples were drawn from 16 healthy individuals after obtaining their signed informed consents. The selected experimental group was age and gender equilibrated. BMI was calculated for each individual enrolled. The inclusion criterion was age (20–60 years). The exclusion criteria were the following: acute or chronic diseases, pregnancy, smoking and taking oral contraception or other drugs. Every volunteer was screened for the viral and bacterial infection markers of blood-transmittable diseases (Syphilis, HIV, Hepatitis B/C). Blood samples from fasting morning participants were collected between 7 am and 9 am, on three separate days within a period of one month [260].

[7] blank: incubation of only PBS buffer with the secondary detection antibody.

[8] This *Methods* section is not exhaustive with respect to experimental materials and methods. Only the necessary information needed to understand the experiments as well as to understand the Results and Discussion section is presented in this section. In addition, as mentioned in Section 1.8.2 (page 22), except for the Mouse study (Section 3.5.8), only results (Chapters 4–8) regarding signal intensity profiles of *healthy* individuals of each experimental study are reported to remain within this work's scope. For completeness reasons, however, sample numbers of non-healthy individuals are listed for each experimental study.

Slovenian healthy study: Quantification of plasma IgM

Serum IgM was quantified using the human IgM quantification sets (Bethyl Labs). EIA (Enzyme Immunoassay) plates included standard, blank and positive control samples (serum pool). Briefly, Corning 96-well EIA plates (Sigma Aldrich) were coated for 1 h with anti-human IgM diluted 1:100 in coating buffer (carbonate-bicarbonate, pH=8) and blocked for 30 min in blocking buffer (TBS, 1% BSA, pH 8). Samples were pre-diluted 1:600 (IgM) in sample diluent (TBS, 1% BSA, 0.05% Tween20, pH 8), serially two-fold diluted across columns and incubated for 3 h. HRP (horseradish peroxidase)-coupled anti-human IgM was diluted 1:5000 in sample diluent and incubated for 2 h. ABTS substrate was incubated for 5 min, and absorbance was measured at 405 nm.

Slovenian healthy study: Serum antibody binding assays

Sera were 1:10 diluted in blocking buffer and incubated automatically on individual microarrays as detailed in Section 3.2.2. Additionally, one microarray was, instead of serum, incubated with blocking buffer (PBS, 1% BSA) (negative control, blank) and two repeated incubations were performed to characterize technological variability. Serum IgM antibody binding was detected with goat-anti-human IgM-Alexa Fluor 647 secondary antibody (20 µg/ml).

3.5.2 Glioma 09 study

Incubations, serum IgM quantification and signal detection were performed by both Juliane Lück and Bodo Steckel. All samples originate from the lab of Christian Schichor (Ludwig-Maximilians-Universität, Munich, Germany). For the study's background, please refer to Lück (PhD thesis) [259].

Glioma 09: Serum samples

Serum samples from primary astrocytoma patients (Table 3.3) were collected prior to any radiotherapy, multimodal treatment (excluding cortisone intake) and surgical tumor resection. Patients were classified into low malignant histological tumor subtypes (*LM*: WHO grade I and II) and high malignant histological tumor subtypes (*HM*: WHO grade III and IV) were included (Table 3.3). The histological grade was confirmed by histopathological analysis. Patients with tumor recurrence or progression, and patients with non-Glioma benign lesions were excluded from the study. Blinded samples (diagnosis not known to the Systems Immunology research group) were also included in this study.

Glioma 09: Quantification of serum IgM

Serum quantification was done as for the SHS (Section 3.5.1) with the exception that sera were diluted 1:500.

Experimentator	Healthy	LM	HM	Serum pool	Blinded
Juliane Lück	17 (17)	7 (7)	31 (31)	4 (4)	0 (0)
Bodo Steckel	1 (0)	7 (6)	2 (1)	0 (0)	9 (9)
Total	18 (17)	14 (13)	33 (32)	4 (4)	9 (9)

Table 3.3: Glioma 09 study: Incubated samples analyzed in the Glioma 09 study. Brackets indicate sample numbers *without* repeats. Juliane Lück used serum pool samples of healthy human individuals as repeats to control for technological variability. The repeats, performed by Bodo Steckel, served mainly for comparing whether signal intensities depend on the experimentator. This was found to be not the case (data not shown).

Glioma 09: Serum antibody binding assays

Incubations were performed automatically using the Tecan HS 4800 hybridization station. Serum was applied in a concentration of 80 µg/ml, diluted in PBS/BSA. IgM-antibody binding was detected using anti-human IgM-Alexa Fluor 647. Isotype-specific secondary antibodies were applied in a concentration of 20 µg/ml. Additionally, one microarray was, instead of serum, incubated with blocking buffer (PBS, 1% BSA) (negative control, blank) and repeated incubations were performed to characterize technological variability (Table 3.3).

3.5.3 Glioma 08 study

The two major differences between the Glioma 08 and the Glioma 09 studies are: (i) fewer samples were used in the Glioma 08 study, (ii) serum samples were diluted 1:10 in the Glioma 08 study whereas for the Glioma 09 study the IgM concentration of incubated samples was set to 0.08 mg/ml.

Incubations and signal detection were performed by Juliane Lück. All samples originate from the lab of Christian Schichor (Ludwig-Maximilians-Universität, Munich, Germany).

Glioma 08: Serum samples

In the Glioma 08 study, healthy controls ($n = 17$), low malignant histological tumor subtypes (*LM*: WHO grade I and II, $n = 8$) and high malignant histological tumor subtypes (*HM*: WHO grade III and IV, $n = 24$) were included.

Glioma 08: Quantification of serum IgM

Please refer to Section 3.5.2 as methods were analogous.

Glioma 08: Serum antibody binding assays

In total, 49 distinct sera (32 patient sera and 17 healthy control sera) were diluted 1:10 in blocking buffer and were manually incubated (Section 3.2.1) on individual peptide microarrays. Additionally, one microarray was, instead of serum, incubated with blocking

buffer (PBS, 1% BSA) (negative control, blank) and two repeated incubations of a healthy control serum were performed to characterize technological variability. Serum IgM antibody binding was detected with goat-anti-human IgM-Alexa Fluor 647 secondary antibody (20 µg/ml).

3.5.4 NephroFIT study

This study was originally designed to study whether sera from kidney transplanted patients who *rejected* the graft could be separated by antibody binding profile from those who *did not reject*. Blood samples were collected thirty days after transplantation. Patients were then monitored for the period of one year in order to follow their rejection behavior. The goal was to detect differentiating patterns with regard to those patients that rejected the transplant and with regard to those who did not reject.

Serum incubations and signal detection were performed by Bodo Steckel. Serum samples were obtained from Nina Babel, Thomas Schachtner and Petra Reinke (Department of Nephrology, Charité, Berlin, Germany).

NephroFIT: Serum samples

Blood samples were collected thirty days after transplantation. Patients were monitored for the period of one year in order to follow their rejection behavior. For this purpose, they were divided into two groups: "Rejection" ($n = 15$), and "No rejection" ($n = 14$).

NephroFIT: Antibody binding assays

The serum-array incubations were performed automatically using the Tecan HS 4800 hybridization station (Section 3.2.2) with serum, which was diluted 1:10 in PBS/BSA (blocking buffer). Additionally, one microarray was, instead of serum, incubated with blocking buffer (PBS, 1% BSA) (negative control, blank) and two repeated incubations with healthy control sera were performed to characterize technological variability. IgM-antibody binding was detected using goat anti-human IgM-Alexa Fluor 647. This secondary antibody was applied in a concentration of 20 mg/ml.

3.5.5 NephroFIT-Pepscan study

In contrast to the NephroFIT study, this study was carried out with peptide arrays from Pepscan and has fewer samples. For further background on platform differences between JPT and Pepscan microarrays, please refer to Section 3.1. Incubations and signal detection was performed by Juliane Lück. Serum samples were obtained from Nina Babel, Thomas Schachtner and Petra Reinke (Department of Nephrology, Charité, Berlin, Germany).

NephroFIT-Pepscan: Serum samples

Blood samples were collected thirty days after transplantation. People were then followed for the period of one year in order to follow their rejection behavior. For this purpose,

they were divided into two groups: “Rejection” ($n = 13$), and “No rejection” ($n = 13$).

NephroFIT-Pepscan: Antibody binding assays

The serum-array incubations were performed manually (Section 3.2.1) with serum, which was diluted 1:10 dilution in PBS (1% BSA). Additionally, one microarray was, instead of serum, incubated with blocking buffer (PBS, 1% BSA) (negative control, blank) and two repeated incubations with healthy control sera were performed to characterize technological variability. IgM-antibody binding was detected using goat anti-human IgM-Alexa Fluor 647. This secondary antibody was applied in a concentration of 20 µg/ml.

3.5.6 NephrOT study

Briefly, this study was to study whether differences in rejection behavior of kidney transplant recipients can also be detected by antibody profiling. Studied rejection behaviors are tabled in Section 3.5.6.

Incubations, plasma IgM quantification and signal detection were performed by Juliane Lück, Rafael Burtet and Andrea Maranhao.

NephrOT: Plasma samples

Plasma samples were obtained from two different centers of the Operational tolerance multicenter study: Rio Grande do Sul (PUC-RS) and of the clinics hospital of medicine school of São Paulo University (HCFMUSP).

Characterization of studied groups:

- *Group OT* (operational tolerance): Individuals with long term stable transplant (> 1 year of transplantation), without use of immunosuppressive drugs or at least 1 year. Number of samples: 5.

- *Group CR* (chronic rejection): Individuals with chronic rejection long time after transplantation (> 1 year of transplantation) (diagnosed by biopsy using histopathologic criteria for the classification of the transplanted kidney [261, 262]). Number of samples: 8.

- *Group ST* (stables): Individuals with long term stable transplant (> 1 year of transplantation) treated with standard doses of immunosuppressants. Number of samples: 8.

- *Group HE* (healthy): Healthy donor for kidney transplantation. Number of samples: 9.

NephrOT: Quantification of plasma IgM

Plasma IgM was quantified using the human IgM quantification set (Bethyl Labs). EIA plates included standard, blank and positive control samples (human reference serum, Lonza). All blood plasma samples and the standard were assayed in duplicate. Briefly, corning 96-well EIA plates (Sigma Aldrich) were coated for 1 h with anti-human IgM diluted 1:100 in coating buffer (carbonate-bicarbonate, pH=8) and blocked for 30 min in blocking buffer. (TBS, 1% BSA, pH 8). Samples were pre-diluted 1:300 in sample diluent (TBS, 1% BSA, 0.05% Tween20, pH 8), serially two-fold diluted across columns and incubated overnight at 4°C. HRP-coupled anti-human IgM was diluted 1:5000 in sample diluent and incubated for 1 h. TMB substrate was incubated for 1 min, the reaction was stopped with 0.18 M sulfuric acid, and absorbance was measured at 450 nm.

NephrOT: Antibody binding assays

The plasma-array incubations were performed automatically using the Tecan HS 4800 hybridization station (Section 3.2.2).

Plasma samples were diluted 1:10 in blocking buffer. IgM-antibody binding was detected using anti-human IgM-Alexa Fluor 647. Secondary antibodies were applied in a concentration of 20 µg/ml. Additionally, one microarray was, instead of serum, incubated with blocking buffer (PBS, 1% BSA) (negative control, blank) and 6 repeated incubations of a healthy control serum were performed to characterize technological variability.

NephrOT: Signal detection

Microarrays were scanned with the Genepix4200AL microarray scanner. The PMT (Photomultiplier Tube) gain was set to 350, the laser power to 20%[9] and the resolution to 10 µm.

3.5.7 NOD study (NS)

Christin Schläwicke performed incubations involving samples of healthy C57BL/6 mice and healthy NOD (non-obese diabetic) mice.

NOD study: Serum samples

Serum was taken from female 5 healthy C57BL/6 and 5 NOD (non-obese diabetic, still healthy), mice. Mice were provided by the Disease Genetics Group (AG Penha-Gonçalves), Instituo Gulbenkian de Ciência, Oeiras, Portugal.

NOD study: Serum antibody binding assays

Serum samples (1:10 diluted in blocking buffer) were manually incubated (Section 3.2.1). Additionally, one microarray was, instead of serum, incubated with blocking buffer (PBS,

[9] PMT gain and laser power were set to different values than for all other studies (Section 3.3) and are therefore reported here.

1% BSA) (negative control, blank). Serum IgM antibody binding was detected with goat-anti-mouse IgM Alexa Fluor 546 secondary antibody (20 µg/ml).

3.5.8 Mouse study (MS)

Serum samples from 15 BALB/c mice bred under specific pathogen-free (SPF) conditions were collected. These mice were infected with *Heligmosomoides bakeri* HB (Figure 3.1) Further serum samples were collected at 10 dpi (days post infection; 15 samples), at 14 dpi (13 samples) and at 18 dpi (15 samples) totaling 58 serum samples (Figure 3.1). The goal of this study was to discriminate the binding patterns of samples obtained from different time points after infection. For an immunological background on murine HB infections, please refer to Section 1.8.2.

In addition to sera from BALB/c mice, also 15 sera from distinct healthy C57BL/6 SPF-mice were analyzed.

Incubations, serum IgM quantification and signal detection were performed by Juliane Lück.

Mouse study: Mice

Male BALB/c mice were bred and maintained under specific pathogen-free (SPF) conditions by the Department of Molecular Parasitology, Humboldt University Berlin (Sebastian Rausch). Infection of mice with HB was carried out by oral gavage with 200 L3 stage larvae in distilled water (Figure 3.1).

C57BL/6 mice were bred and maintained under specific pathogen-free (SPF) conditions and were provided by the lab of Marc Ehlers (former Research Group "Tolerance und Autoimmunity", DRFZ, Berlin, Germany) and by the lab of Simon Fillatreau ("Immunregulation", DRFZ, Berlin, Germany).

Mouse study: Serum samples

Mice of both groups (healthy C57BL/6 [$n = 15$] and BALB/c mice [$n = 58$]), Section 3.5.8) were narcotized and bled either by cardiac or retro-orbital puncture at the age of 8 weeks.

As to BALB/c mice, blood samples were collected from healthy SPF-BALB/c mice ($n = 15$), which were subsequently infected with HB. Blood samples were collected at three time points post infection (dpi): at 10 dpi ($n = 15$), 14 dpi ($n = 13$) and 18 dpi ($n = 15$).

The blood was allowed to clot at room temperature, centrifuged and the supernatant was stored at $-20°C$.

Mouse study: Quantification of serum IgM

Serum IgM quantification was performed as described for the NephrOT study (Section 3.5.6) with the exception that sera were prediluted 1:500. The IgM concentration was only measured—due to technical complications—for 32 samples (*HE*: 12, *AP*: 15, *CP*: 5).

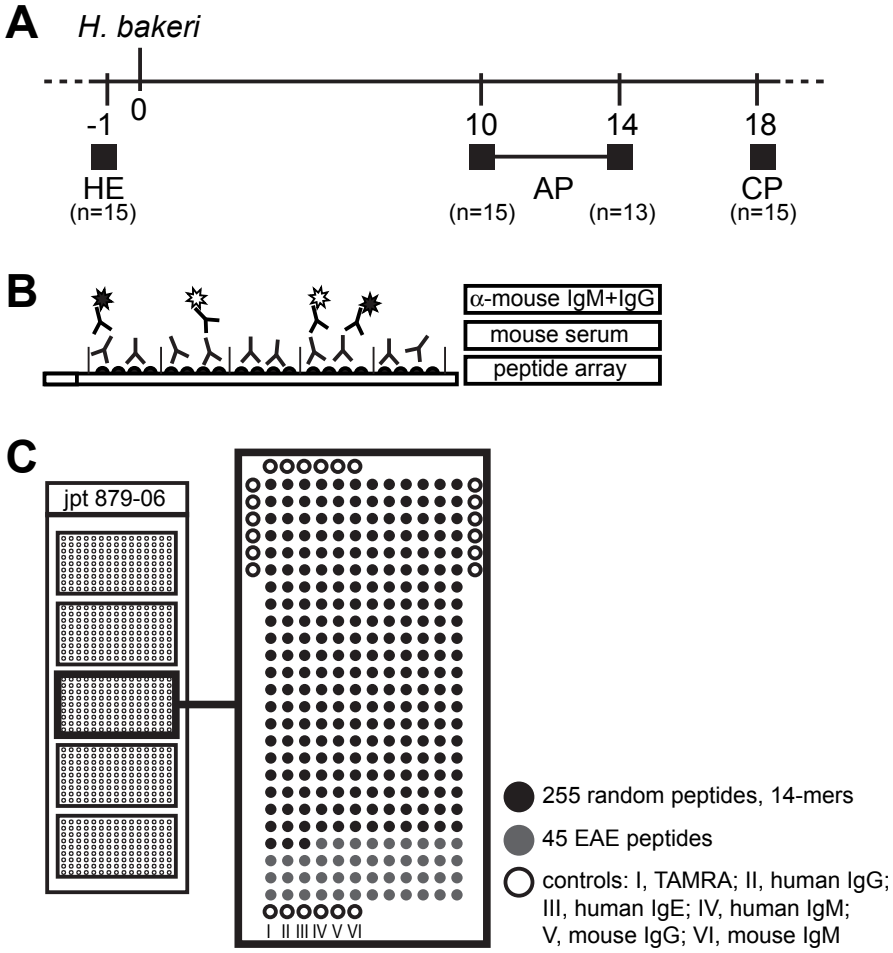


Figure 3.1: General experimental setup. (A) *Heligmosomoides bakeri* (HB)-infection of mice was carried out by oral gavage with 200 L3 stage larvae in distilled water. Sera were prepared from blood samples collected prior to infection (*HE*), during acute infection (*AP*) or early chronic infection (*CP*). (B) Sera were probed on peptide microarrays and afterwards bound serum antibodies were detected with fluorescence- labeled anti-mouse IgM. (C) Microarray slides covered five identical sub-arrays, each including 300 peptide (255 random-sequence peptides [$J^{255}_{14\text{-mer}}$] and 45 EAE-peptides) spots and four replicates of six internal controls. Only the 255 random-sequence peptides were analyzed in this study. This Figure was conceived by Nicole Wittenbrink.

Mouse study: Antibody binding assays

Sera were manually incubated as detailed in Section 3.2.1. Each sub-array was incubated with a different serum. This is in contrast to all other studies (Section 3.3.2). Additionally, one sub-array was, instead of serum, incubated with blocking buffer (PBS, 1% BSA, negative control, blank). Furthermore, three arrays of a different batch but analogous design were incubated with the same serum of a healthy mouse in order to characterize technological variability. Serum antibody binding was detected with polyclonal goat anti-mouse IgM-Alexa Fluor 546.

3.5.9 Monoclonal antibodies

The 13 human monoclonal antibodies were kindly provided by the group of Hedda Wardemann (Max Planck Institute for Infection Biology, Berlin, Germany). Ten different Ig gene sequences of IgG^+ memory B cells from 2 healthy human donors, PN and VB, (PN115, PN138, PN16, PN89, VB1, VB142, VB161, VB176, VB18, VB4) [76] and three further ones from 3 other human donors ED38 [263], eiJB40 and mGO53 [74] were expressed as detailed in Tiller and colleagues [264].

Monoclonal antibodies were manually incubated analogously to sera from the Mouse study (Section 3.5.8) with a concentration of 10 mg/ml on sub-arrays displaying the $J^{255}_{14\text{-mer}}$ library (Section 3.1.1).

3.6 Simulation of antibody-peptide reactivity data

3.6.1 Simulation of signal intensities

Peptides ($\vec{p}^{\,i}$) and antibody binding sites ($\vec{a}^{\,k}$) were modeled as strings. Binding strengths between antibodies and the various amino acid residues of a peptide, referred to as assigned AAWS $\vec{h}$, were sampled from the uniform distribution on the closed interval $[0, 1]$. A binding site on an antibody $\vec{a}^{\,k}$ was simulated in a similar fashion with a random number from the closed interval $[-1, 1]$ (i.i.d.) for every sequential position and scaled such that $(\vec{a}^{\,k})^{\mathrm{T}}\vec{a}^{\,k} = 1$. The binding association between peptide $\vec{p}^{\,i}$ and antibody $\vec{a}^{\,k}$ was calculated by $y_{i,k} = (\vec{a}^{\,k})^{\mathrm{T}}\vec{p}^{\,i}$ (Section 4.2).

Based on the interpretation of the binding association as being negatively linearly proportional to the standard Gibbs free energy change of reaction, $\Delta_r G^\circ$, the binding affinity $K_{i,k}$, the thermodynamic equilibrium association constant for antibody k binding peptide i, is defined as shown in Equation 3.2.

$$K_{i,k} = \exp\left(-\frac{\Delta_r G^\circ}{RT}\right) = \exp\left(\frac{\beta_0 + \beta_1 y_{i,k}}{RT}\right) \tag{3.2}$$

Similar to a bit string model approach published by Rosenwald and colleagues [265], the calculation of $K_{i,k}$ assumes additivity in free energy of binding[10], an assumption that is supported by experimental results [266, 267]. The measured signal intensity is assumed to be proportional to the ratio of bound-to-total surface of the peptide spot, S_i. An expression for this quantity, based on the law of mass action, can be obtained from classical Langmuir adsorption theory [268] resulting in the following Equation with $R = 8.314472$, $T = 273.15 + 25$, $\beta_0 = 0$ and $\beta_1 = RT$.

$$S_i = \frac{\sum_{k=1}^{n_{\mathrm{Ab}}} [\mathrm{Ab}]_k K_{i,k}}{1 + \sum_{k=1}^{n_{\mathrm{Ab}}} [\mathrm{Ab}]_k K_{i,k}}, \tag{3.3}$$

where $[\mathrm{Ab}]_k$ is the concentration of antibody k with $\sum_{k=1}^{n_{\mathrm{Ab}}} [\mathrm{Ab}]_k = [\mathrm{Ab}]_{\mathrm{Total}}$. The total

[10] This thermodynamic insight was contributed by the author of this thesis.

antibody concentration $[\mathrm{Ab}]_{\mathrm{Total}}$ is set to 1 unless mentioned otherwise[11]. Equation 4.1 has been conceived by Henning Redestig, Johannes Schuchhardt and Michal Or-Guil and was first published in Greiff and colleagues [104].

At last, signal intensities were normalized (log-transformed, mean-centered, and scaled to unit variance) if not stated otherwise.

3.6.2 Introduction of Gaussian noise into simulated signal intensities

If Gaussian noise ($\mathcal{N}(\mu, \sigma)$) was introduced into simulated signal intensities, the noise term—an i.i.d. generated vector of the length of the antibody sequence—was multiplicatively introduced before logarithmic transformation of the data.

3.6.3 Simulation of correlated antibody repertoires

Correlated antibody repertoires were simulated as shown in Figure 3.2. Briefly, correlated antibody repertoires were simulated by adding Gaussian noise $\overrightarrow{\mathcal{N}(\mu = 0, \sigma = x(j))}^{i}$ ($i \in \{1, 2, \ldots, 10^4\}$) to an antibody sequence $\vec{a}$. Repeating this 10^4 times with the same antibody sequence $\vec{a}$ yields a set of 10^4 sequences, $\{\vec{a}^{\,i}_{\mathrm{noise}}\}_{i \in \{1,2,\ldots,10^4\}}$, wherein the set's internal correlation structure depends on the amount of Gaussian noise added to an antibody sequence (Table 3.4). The amount of noise added to an antibody sequence is a function of $x(j)$, where $x(j) \in [0, 10], j \in \{1, 2, \ldots, 10\}$. All antibody sequences were normalized to $(\vec{a}^{\,i}_{\mathrm{noise}})^{\mathrm{T}} \vec{a}^{\,i}_{\mathrm{noise}} = 1$ after Gaussian noise was added.

Median correlation coefficient (r)	1.00	1.00	0.86	0.61	0.27	0.14	0.08	0.05	0.00	0.00
Gaussian noise ($\mathcal{N}(\mu = 0, \sigma = x)$)	0.00	0.01	0.10	0.20	0.40	0.60	0.80	1.00	5.00	10.00

Table 3.4: Assessment of the dependence of the median correlation of correlated antibody repertoires on the level of Gaussian noise introduced into antibody sequences (Figure 3.2). With increasing Gaussian noise, the degree of decorrelation of generated antibody repertoires increases. The medians of the median pairwise Pearson correlation coefficients of the 40 generated repertoires (Section 6.5, Figure 6.11) are tabled (1st row) in function of Gaussian noise (2nd row). No dependence on antibody strength was found (data not shown).

Decorrelation describes an increase in the standard deviation ($\sigma = x(j)$) of the added Gaussian noise accompanied with a decrease in pairwise Pearson correlation between antibody sequences of a given repertoire (Table 3.4).

3.7 Partial least squares regression

3.7.1 Estimation of AAWS with PLSR

Let $\vec{S}$ be a vector, which itemizes the measured or simulated signal intensities and $\mathbf{X}$ the *amino acid composition matrix* (AACM) where $x_{i,q}$ is number of amino acid residues of

[11] The importance of a constant total antibody concentration for the simulation of signal intensities was contributed by the author of this thesis.

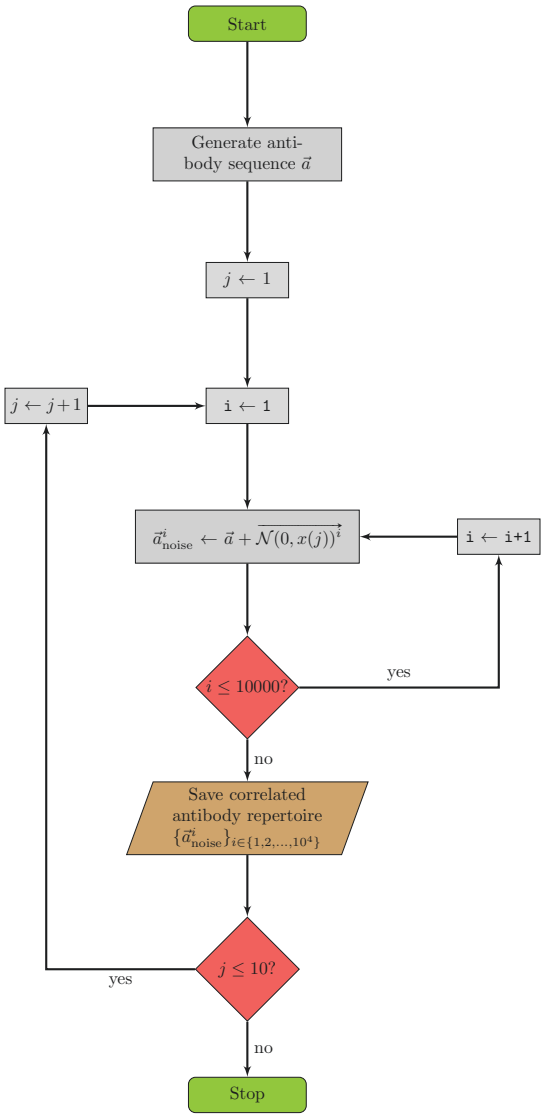


Figure 3.2: Flowchart of the generation of correlated antibody repertoires. For further information, see main text (Section 3.6.3).

type q in peptide $\vec{p}^{\,i}$ (Equation 4.8).

The response variable $\vec{S}$ and the matrix of predictor variables $\mathbf{X}$ cannot be correlated directly with each other due to non-congruent dimensionality. Thus, a to be determined vector $\vec{w}$, which relates the predictor matrix $\mathbf{X}$ to the both log-transformed and scaled ($\mu = 0, \sigma = 1$) response variable $\vec{S}$, is needed. If $\vec{S}$ is of dimension $m \times 1$ and $\mathbf{X}$ of dimension $m \times n$, then, according to Equation 4.8, $\vec{w}$ must be of dimension $n \times 1$. If the variance of the residual vector $\vec{\epsilon}$ in Equation 4.8 is small, one would expect the regression coefficients composing $\vec{w}$ to vary sensitively with any changes in $\vec{S}$, thus displaying the

importance of every amino acid for signal intensity generation.

In theory, $\vec{w}$ can be estimated by ordinary least squares regression (OLR), i.e. $\hat{\vec{w}} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\vec{S}$, however, unless the peptide library was designed for this purpose, $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ may not be invertible and OLR fails. In this case, it is still possible to obtain a robust estimate of $\vec{w}$ by means of Partial Least Squares Regression [191, 269, 270].

PLSR is used to find the fundamental relations between two matrices (or one vector [$\vec{S}$] and one matrix [$\mathbf{X}$]). It is a latent variable approach to modeling the covariance structures in these two spaces. A PLSR model tries to find the multidimensional direction in the $\mathbf{X}$-space that explains the maximum (multidimensional) variance direction in the $\vec{S}$-space. PLSR is particularly suited when the matrix of predictors has more variables than observations, and when there is collinearity in $\mathbf{X}$. An introduction to PLSR can be found in Wold and colleagues [270]. More extensive mathematical foundations, especially regarding the latent variables, are presented in the Appendix (Section A.1).

All calculations involving PLSR were carried out with the `pls` R package (Version 2.10) [271, 272].

3.7.2 PLSR model diagnostics

In order to assess the quality of the built PLSR model, Q^2 was used as a measure for assessing the predictive performance. The predictive performance is defined as:

$$Q^2 = 1 - \frac{\sum(\hat{S}_{\text{Left out}} - S_{\text{Left out}})^2}{\sum S^2_{\text{Left out}}}. \tag{3.4}$$

The vector $\vec{S}_{\text{Left out}}$ is the left-out test data set, the signal intensity of which is predicted ($\hat{\vec{S}}_{\text{Left out}}$) from the remaining training data set. The left-out test data represented randomly chosen 10% of the total data set. The predictive performance was assessed for a varying number of latent components (Section A.1). The number of latent components which minimized the cross-validated prediction error, that is, which maximized the predictive performance (Q^2), was chosen.

3.8 Unsupervised and supervised machine learning methods

3.8.1 Principal component analysis

Principal component analysis (PCA) is an unsupervised[12] technique in pattern recognition. It transforms the input data set in such way that it may be represented by a reduced number of "effective features", yet retains most of the intrinsic information content of the original data. Simply truncating a vector $\vec{x}$ would result in a mean-square error equal to the sum of the variances of the elements eliminated from $\vec{x}$. PCA, however, provides a linear invertible transformation of the data space. This yields a truncation of $\vec{x}$, which is optimal in mean-square-error sense; PCA projects the data onto a new coordinate system such that the greatest variance by any projection of the data lies on the first

[12] Unsupervised methods are mostly used for pattern recognition and dimension reduction purposes [273].

coordinate (called the first principal component), the second greatest variance on the second coordinate, and so forth. PCA is *not* optimized for class separability [273, 274].

Principal component analysis was done with the `pca`-function of the `pcaMethods` R package (Version 1.32) [272, 275].

3.8.2 Support vector machines

Support vector machines are binary[13] learning machines, which are mostly used in pattern-classification problems. They are categorized as large margin classifiers. Given a training sample, the support vector machine constructs a hyperplane as the decision surface maximizing the margin of separation between positive and negative examples.

A notion that is central to the development of the SVM learning algorithm is the inner-product kernel between a "support vector" $\vec{x}_i$ and a vector $\vec{x}$ drawn from the input data space. The support vectors consist of a small subset of data points extracted by the learning algorithm from the training sample [273, 274, 277].

The optimization problem is normally solved in its dual form with the classification rule being:

$$f(\mathbf{x}) = \operatorname{sgn}\left(\sum_{i=1}^{m} \alpha_i y_i k(\vec{x}_i, \vec{x}) + b\right),$$

where m is the number of training samples, b is the bias, a_i are the Lagrange multipliers, $y_i \in \{-1, 1\}$ is the label vector and $k(\vec{x}, \vec{x}_i)$ is the kernel function. Those vectors for which the Lagrange multipliers α_i are non-zero are called support vectors. They are either found on the classification margin or within the margin.

P-SVM

Support vector machine analysis was carried out with an R implementation of the Potential Support Vector Machine (P-SVM) [278], which facilitates linear classification of high-dimensional data with a built-in feature selection. The classification performance was measured using nested leave-one-out cross-validation (LOOCV), where feature selection and hyperparameter selection were performed in the inner cross-validation loop independently of the test sample of the outer cross-validation loop. The inner loop was used to determine the combination of parameters that gives the best classification performance: the cost parameter c was varied from 1 to 17 in 5 equally spaced steps and the regularization parameter ε was chosen as 2^i with $i = -3, -2, \ldots, 3, 4$. In order to obtain compact models that only use a small set of features, all parameter combinations in the inner cross-validation loop for which more than three models exceeded an upper limit of six selected features were rejected. The goal criterion of classification performance was balanced accuracy (BACC, Equation 3.7). A flowchart of the used algorithm is given in Figure 3.3.

For determining whether selected peptides were unique in their classification accuracy, unique peptides—determined across folds by subproblem (Table S.1)—were removed

[13] Multi-class support vector machine approaches also exist [276].

from the respective subproblem's data set. Afterwards, the cross-validation procedure was carried out for the remaining peptides as described above (Figure 3.3).

All P-SVM computations were done with the `psvm` R package[14] (Version 0.06).

P-SVM: Definition of balanced accuracy, specificity and sensitivity

$$\text{Sensitivity} = \frac{\text{number of true positives}}{\text{number of true positives} + \text{number of false negatives}} \tag{3.5}$$

$$\text{Specificity} = \frac{\text{number of true negatives}}{\text{number of true negatives} + \text{number of false positives}} \tag{3.6}$$

$$\text{Balanced accuracy (BACC)} = \frac{1}{2} \times (\text{Sensitivity} + \text{Specificity}) \tag{3.7}$$

P-SVM: Permutation testing

Since, throughout this thesis, the number of peptides (features) is high compared to the number of samples, correlations between features and the target may occur by chance. In order to quantify such random effects, permutation testing was performed: for 1000 independent random shuffles of the label vector LOOCV[15] was performed. Then, the proportion of outcomes in which the balanced accuracy was at least as good as for the original unshuffled label vector is an unbiased estimate of the true p-value, i. e. the probability to achieve a balanced accuracy at least as good as the observed one under the null hypothesis that labels and input data are independent.

3.9 Statistical analysis

3.9.1 Correlation coefficients

Association between variables was assessed by Pearson correlation (r_{Pearson}) unless stated otherwise, in which case the Spearman-rank correlation (r_{Spearman}) coefficient was used. The Pearson correlation coefficient is a measure of the correlation (linear dependence) between two variables X and Y, $r_{\text{Pearson}}, r_{\text{Spearman}} \in [-1, 1]$.

Pearson's correlation coefficient between two variables (X, Y) is defined as the covariance of the two variables divided by the product of their standard deviations:

$$r_{\text{Pearson}} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}}.$$

[14] This package has not yet been published.

[15] Permutation testing was performed using the best precomputed parameter combinations (ε, c) determined by P-SVM *nested* cross-validation for the non-shuffled data set. This was done to limit computational load.

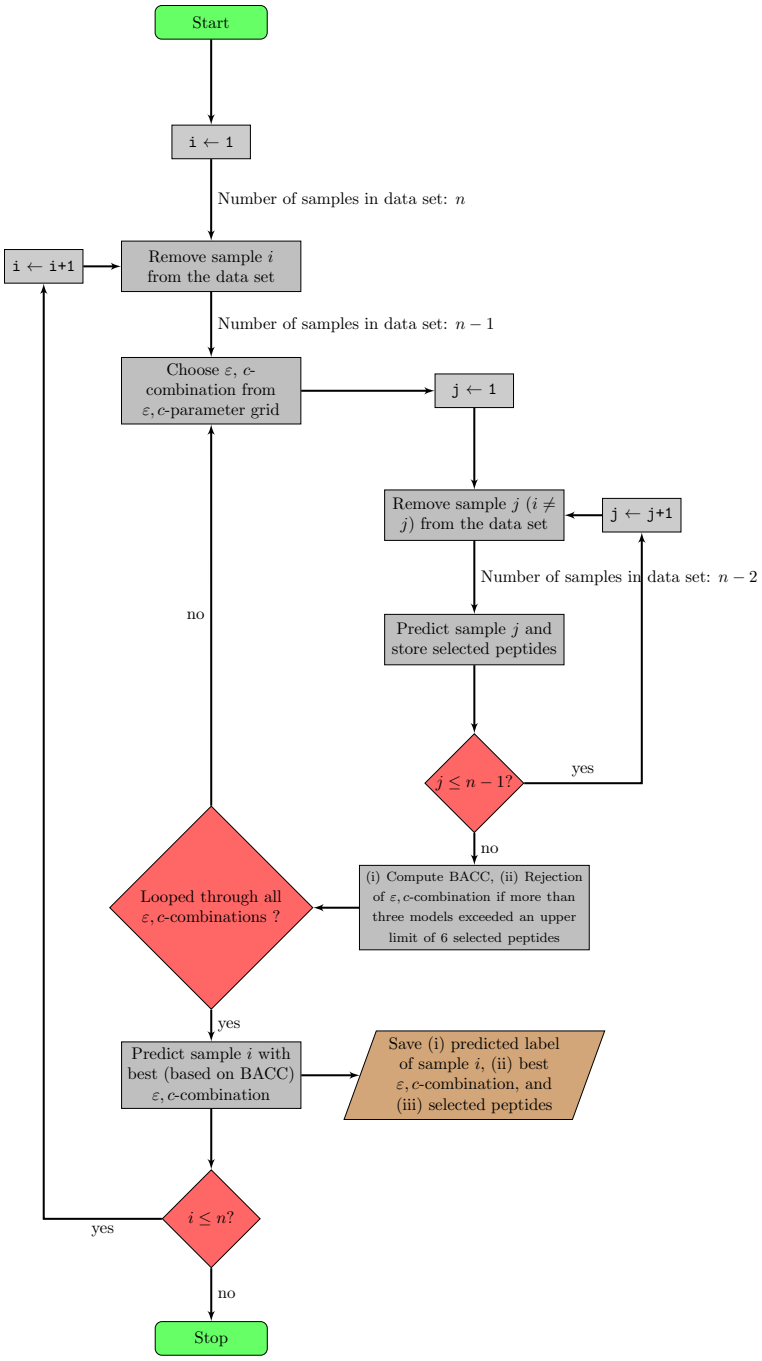


Figure 3.3: Flowchart of the P-SVM algorithm. For further information, see main text (Section 3.8.2).

Correlation coefficients equaling to 1 or -1 correspond to data points lying exactly on a line.

In contrast to the Pearson-correlation coefficient, the Spearmen correlation coefficient is a non-parametric measure of statistical dependence between two variables. It is defined as the Pearson correlation coefficient between the ranked variables. For a sample of size n, X_i, Y_i are converted to ranks x_i, y_i, and r_{Spearman} is computed from these:

$$r_{\text{Spearman}} = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}}.$$

The Spearman correlation coefficient assesses how well the relationship between two variables can be described using a monotonic function. If there are no repeated data values, a perfect Spearman correlation of $+1$ or -1 occurs when each of the variables is a perfect monotone function of the other. Tied values are assigned a rank equal to the average of their positions in the ascending order of the values [279, 280].

3.9.2 Hierarchical clustering

For hierarchical clustering of signal intensity profiles or AAWS, correlation coefficients of the data type in question were Pearson/Spearman-correlated in a pairwise fashion thus yielding a correlation matrix. The dimension of this matrix is equal to the number of pairwise-correlated samples. These coefficients were then transformed into Euclidean distances by the `dist()`[16] function in R thus returning a distance matrix. The "complete-linkage" clustering algorithm performed by the R function `hclust()`[17] proceeds by initially assigning each object to a cluster. It then proceeds iteratively, at each stage joining the two most similar clusters, continuing until there is just a single cluster. At each stage distances between clusters are recomputed by the Lance-Williams dissimilarity update formula [281]. When joining a new object (signal intensity profile or AAWS) to an existing node, the distance of this new object from the node is the largest distance found from the new object to all objects contained within that node. This method provides the greatest separation of clusters when compared to other types of clustering such as "single-linkage" or "average-linkage clustering" [282]. Hierarchical clustering creates a dendrogram, which is passed on to the `heatmap.2()`[18] R function creating a heatmap of the clustered correlation matrix.

Heatmaps are false color images with an added dendrogram. Ordering of the rows and columns is imposed by the restrictions of the dendrogram. In this thesis, a histogram displaying the distribution of Pearson/Spearman correlation coefficients is shown in the upper left corner of the heatmap.

3.9.3 Significance testing

The two-sided, non-paired Wilcoxon rank-sum test was used to test—if not mentioned otherwise—the difference between two samples of independent observations. P-values were regarded as significant when $p < 0.05$.

[16] `stats` R package, Version 2.12.1, [272].
[17] `stats` R package, Version 2.12.1, [272].
[18] `gplots` R package, Version 2.10.1, [272, 283].

4 A minimal model of antibody-peptide binding: mathematical analysis and simulations

Parts of this Chapter were recently published [104].

4.1 Preliminary definitions

Definition 4.1.1. An antibody repertoire is the set of all unique antibody strings[1] $\vec{a}^k$ in a given entity. An antibody repertoire is thus defined as $\mathrm{A}_j = \{\vec{a}^k | k \in \mathbb{N}_{\geq 1}\}$. The cardinality of the antibody repertoire A_j ($\#\mathrm{A}_j$) is denoted as $n_{\mathrm{Ab}_{\mathrm{A}_j}}$.

Definition 4.1.2. An antibody mixture is the entirety of antibody strings in a given entity and defined as an indexed family $\mathrm{AM}_j = (\vec{a}^j)_{j \in J}$. The set of antibody concentrations in an antibody mixture is $\mathrm{C}_j = \{[\mathrm{Ab}]_k | k \in \mathbb{N}_{\geq 1}\}$, where $[\mathrm{Ab}]_k$ is the concentration of antibody $\vec{a}^k$. The sum of all antibody concentrations $[\mathrm{Ab}]_k$ in a given mixture is the total antibody concentration $\sum_k [\mathrm{Ab}]_k = [\mathrm{Ab}]_{\mathrm{Total}}$.

4.2 A minimal model of antibody-peptide binding

In the following, a model that simulates binding between peptides and antibodies is presented. In this model, the binding affinity of simulated monoclonal antibodies depends *non-linearly* on amino acid positions in the peptide sequences (Equation 4.1). The proposed model is similar to bit string models [68, 284–286] in that it uses vectors as representations of peptides and antibodies.

The peptide string is represented by unique real numbers taken from a vector of assigned amino acid-associated weights (AAWS), denoted $\vec{h}$, the twenty components of which were drawn from a uniform distribution on the closed interval $[0, 1]$. A peptide $\vec{p^i}$ of l amino acids is thus represented by a vector of l numbers drawn from $\vec{h}$.

Definition 4.2.1. A peptide $\vec{p}^i$ is defined as an oriented string of amino acids $p^i_k \in \mathrm{AA}$, where AA is the alphabet amino acids are drawn from. An assortment of peptides forming an indexed family P is called a peptide library $P = (\vec{p}^j)_{j \in J}$.

An antibody binding site is represented by a vector $\vec{a}^k$ of length l. The binding strength of each position is given by a number between -1 and 1 that is drawn randomly from a

[1] Hereafter, the terms "string" and "sequence" will be used interchangeably.

uniform distribution (i.i.d.) and is scaled such that $(\vec{a}^k)^{\mathrm{T}}\vec{a}^k = 1$. The binding association between peptide $\vec{p}^i$ and antibody $\vec{a}^k$ is computed as the dot product of the two vectors, $y_{i,k} = (\vec{a}^k)^{\mathrm{T}}\vec{p}^i$. Thus, the binding association $y_{i,k}$ depends explicitly on an amino acid's position in a given peptide sequence.

Antibody and peptide sequences were modeled such that, unilaterally, positive values in the antibody sequence increase the binding association $y_{i,k}$, whereas negative values in the antibody sequence cause a decrease of $y_{i,k}$. If both assigned AAWS $\vec{h}$ and components of antibody sequences could assume negative contributions to binding, then, in contrast to biochemical intuition, negative values at a given position i of both peptide and antibody binding site would yield a positive contribution to signal intensity and, therefore, would not be different from double positive entries.

Based on the law of mass action and classical Langmuir adsorption theory [268], an expression for the peptide signal intensity yielded by a given antibody mixture is

$$S_i = \frac{\sum_{k=1}^{n_{\mathrm{Ab}}}[\mathrm{Ab}]_k K_{i,k}}{1 + \sum_{k=1}^{n_{\mathrm{Ab}}}[\mathrm{Ab}]_k K_{i,k}} = \frac{s_i}{1 + s_i}. \tag{4.1}$$

The thermodynamic equilibrium association constant for antibody k binding to peptide i is defined as $K_{i,k} = \exp\left(-\frac{\Delta_r G^\circ}{RT}\right)$ with $\Delta_r G^\circ = -(\beta_0 + \beta_1 y_{i,k})$. For further details on parameter values, please consult Section 3.6. Equation 4.1 has been conceived by Henning Redestig, Johannes Schuchhardt and Michal Or-Guil and was first published in Greiff and colleagues [104].

Definition 4.2.2. A signal intensity *profile* $(\vec{S})$ is defined as a vector of signal intensities (S_j) for an arbitrary antibody mixture AM_k and a given peptide library P. S_j is the signal intensity of peptide $\vec{p}^j$. The terms signal intensity profile or antibody binding profile are used interchangeably to denote $\vec{S}$.

Simulated signal intensity profiles were normalized (log-transformed, mean-centered, and scaled to unit variance) if not stated otherwise.

Non-normalized simulated signal intensity profiles of monoclonal antibodies ($n_{\mathrm{Ab}} = 1$) show a lognormal shape (Figure 6.3), whereas those of highly diverse antibody mixtures ($n_{\mathrm{Ab}} = 10000$) have a Gaussian shape (Figure 6.2).

4.3 Mathematical and in silico analysis of the minimal model of antibody-peptide binding

4.3.1 Randomly generated, highly diverse antibody mixtures render the signal intensity profile exclusively dependent on peptide amino acid composition

In the following, the described minimal model (Section 4.2) is mathematically analyzed[2] with respect to the impact of antibody diversity on peptide signal intensity (S_i). It is

[2]The mathematical analysis in Section 4.3.1 was performed by Johannes Schuchhardt and Michal Or-Guil.

assumed, without loss of generality, that $[\mathrm{Ab}]_{\mathrm{Total}} = 1$ and all antibodies have equal concentration ($[\mathrm{Ab}]_k = 1/n_{\mathrm{Ab}}$). Setting $\beta_0 = 0$ and $\beta_1 = RT$, s_i gives:

$$s_i = \frac{1}{n_{\mathrm{Ab}}} \sum_{k=1}^{n_{\mathrm{Ab}}} \exp\left(y_k\right) \tag{4.2}$$

$$s_i = \frac{1}{n_{\mathrm{Ab}}} \sum_{k=1}^{n_{\mathrm{Ab}}} \exp\left((\vec{a}^k)^{\mathrm{T}} \vec{p}^i\right) \tag{4.3}$$

$$s_i = \frac{1}{n_{\mathrm{Ab}}} \sum_{k=1}^{n_{\mathrm{Ab}}} \exp\left(\sum_{z=1}^{l} a_z^k p_z^i\right) \tag{4.4}$$

For this Section's mathematical analysis only, the binding strength at each position of an antibody binding site is given by a number that is drawn randomly from a *Gaussian* distribution. The change from uniform (Section 4.2) to Gaussian distribution greatly simplifies the following derivation.

According to the central limit theorem (CLT) and assuming a mixture for which the number of different randomly generated antibodies tends to infinity ($n_{\mathrm{Ab}} \to \infty$), one can proceed with the mean of the lognormal distribution ($\exp^{(\mu_z + \sigma_z^2/2)}$) for each peptide position z, where $\mu_z = \mu \times p_z^i$ and $\sigma_z = \sigma \times p_z^i$ (μ and σ being the mean and standard deviation used to generate the antibody sequences).

$$s_i = \prod_{z=1}^{l} \exp\left(\mu_z + {}^{\sigma_z^2}/_2\right) \tag{4.5}$$

$$s_i = \exp\left(\sum_{z=1}^{l} \mu_z + {}^{\sigma_z^2}/_2\right) \tag{4.6}$$

For the special case, where $\vec{a}^k$ are modeled with $\mathcal{N}(0,1)$, $\mu_z = 0$ and $\sigma_z^2 = \sigma^2 \times (p_z^i)^2 = \left(\langle (a_z^k)^2 \rangle - \langle a_z^k \rangle^2\right) \times (p_z^i)^2 = (1-0) \times (p_z^i)^2$. Thus, s_i simplifies to:

$$s_i = \exp\left(\sum_{z=1}^{l} \frac{(p_z^i)^2}{2}\right) \tag{4.7}$$

Therefore, for highly diverse and randomly generated antibody mixtures simulated signal intensities (S_i) depend *exclusively* on peptide amino acid composition (Equation 4.7). The order of the amino acids in a peptide sequence $\vec{p}^i$ is rendered unimportant for the signal intensity S_i.

4.3.2 Building a regression model based exclusively on peptide amino acid composition to predict signal intensity profiles

Equation 4.7 indicates that the signal intensity (S_i) simulated with highly diverse and identically and independently distributed (i.i.d.) antibody sequences is independent of

any specific antibody composition as long as antibody diversity is high. In particular, the signal intensity solely depends on peptide amino acids (p_z^i). The simplest ansatz to model the sole dependence of signal intensity profiles $\vec{S}$ ($\#P \times 1$) on peptide amino acid composition would be the following linear regression model.

$$\vec{S} = \mathbf{X}\vec{w} + \vec{\varepsilon}, \tag{4.8}$$

where $\mathbf{X}$ ($\#P \times \#\text{AA}$) is the amino acid composition matrix, $\vec{w}$ ($\#\text{AA} \times 1$) the amino acid-associated weights (AAWS), and $\vec{\varepsilon}$ ($\#P \times 1$) the residuals capturing the part of $\vec{S}$ which cannot be explained by the amino acid composition alone. The $\mathbf{X}$ matrix is formed by counting the occurrences of each of the elements of the used amino acid alphabet (AA) in each peptide which results in a matrix with $\#P$ rows and #AA columns. Importantly, $\mathbf{X}$ does not contain any information about the position of an amino acid in a given peptide sequence. The AAWS vector $\vec{w}$ indicates the contribution of every amino acid to the measured signal intensity. Equation 4.8 has been conceived by Henning Redestig, Johannes Schuchhardt and Michal Or-Guil and was first published in Greiff and colleagues [104].

Once the vector $\vec{w}$ has been estimated, the regression model is used to predict measured signal intensities based solely on the peptides' amino acid composition. The regression model's predictive performance (Q^2) is determined by 10-fold cross-validation (Section 3.7.2). The predictive performance equals 1 for perfect predictions ($\vec{\epsilon} \to 0$) and is close to zero for poor predictions.

AAWS and residuals are estimated by partial least squares regression (PLSR, Sections 3.7 and A.1). Prior to any PLSR, signal intensity vectors $\vec{S}$ are log-transformed, centered to zero and set to unit variance.

4.3.3 Simulations show that the peptide amino acid composition-based prediction of signal intensity profiles improves with increasing antibody diversity

In order to test the above statistical ansatz (Equation 4.8), first, signal intensities for 100 antibodies ($n_{\text{Ab}} = 100$) binding to a peptide library of 255 14-mers were simulated. The peptide library used in the simulation determines the amino acid composition matrix $\mathbf{X}_{\text{sim}}$. Simulated intensities $\vec{S}_{\text{sim}}$ (Figure 4.1A) and respective weights $\vec{w}_{\text{sim}}$ (Figure 4.1B) were estimated using the linear regression model $\hat{\vec{S}}_{\text{sim}} = \mathbf{X}_{\text{sim}}\vec{w}_{\text{sim}}$. The prediction of simulated signal intensities yielded a predictive performance (Q^2) of 0.53, and the Pearson correlation between assigned $\vec{h}$ and estimated AAWS $\vec{w}_{\text{sim}}$ was found to be $r = 0.90$ (Figure 4.1B), which indicates a very good recovery of $\vec{h}$.

Second, signal intensity profiles were simulated with different numbers of antibody variants. The results of the simulation framework are in accord with the mathematical analysis in that predictive performance nears perfection ($Q^2 \to 1$) with a growing antibody diversity ($n_{\text{Ab}} \geq 10000$ antibodies, Figure 4.2A)[3]. Correspondingly, (i) the

[3]This simulation result (Figure 4.2) is independent of the distribution (Gaussian, uniform) used for generating antibody sequences as long as the used values are small. In fact, not only components of

pairwise correlation of computed AAWS ($\vec{w}^{i}_{\text{sim}}$), (ii) the correlation of estimated AAWS with assigned AAWS $\vec{h}$ as well as (iii) the pairwise correlation of signal intensity profiles ($\vec{S}_{\text{sim}}$) near perfection ($r = 1$) with growing antibody diversity (Figures 4.2B–D)[4]. Of note, the correlation of estimated AAWS with assigned AAWS is already high for mixtures of low antibody diversity ($n_{\text{Ab}} = 16$) and low predictive performance (Figures 4.2A and 4.2C).

Thus, there exist antibody mixtures which yield predictive performance values near perfection. Their signal intensity profiles depend almost exclusively on assigned AAWS $\vec{h}$. These mixtures are hereafter called *unbiased*. They are defined by the following properties: (i) the number of different antibody sequences (n_{Ab}) tends to infinity ($n_{\text{Ab}} \to \infty$) and (ii) antibody sequences are generated in a random (i.i.d.) fashion. Section 4.3.4 shows that property (i) can be generalized by not only requiring the overall antibody diversity to be high but also demanding that the diversity of antibodies tends to infinity ($n_{\text{Ab}_{[\text{Ab}]_k}} \to \infty$) for any antibody concentration $[\text{Ab}]_k \in \text{C}_j$. Section 4.3.4 shows that a violation of this property may decrease the regression model's predictive performance.

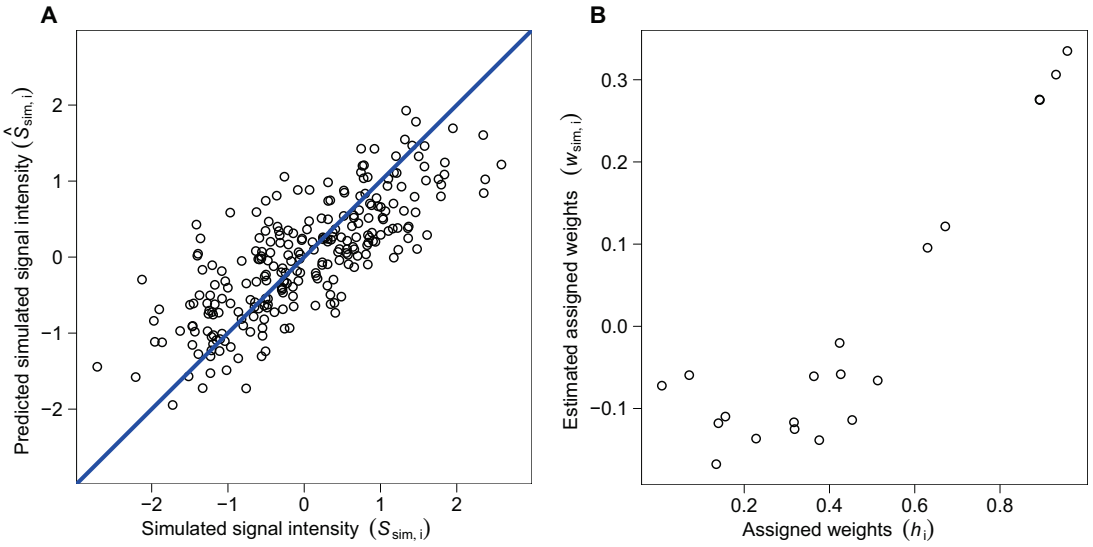


Figure 4.1: Simulated signal intensities and assigned amino acid-associated weights are recovered by an amino acid composition-based regression model. (A) Normalized simulated signal intensities ($\vec{S}_{\text{sim}}$) obtained by amino acid position-dependent simulation of the binding of 100 antibodies to a random peptide library of 255 14-mers were predicted using the regression model given by Equation 4.8. This regression model takes into account only the amino acid composition of the simulated peptide sequences ($\mathbf{X}_{\text{sim}}$). Predictive performance: $Q^2 = 0.53$. Simulated signal intensities were computed using Equation 4.1. (B) Equation 4.8 was used to estimate ($\vec{w}_{\text{sim}}$), which are shown against the assigned AAWS ($\vec{h}$). Pearson correlation coefficient: $r = 0.90$.

antibody sequences but also those of assigned AAWS have to be small ($\approx -1 \leq a^k_i, h_i \leq 1$) so that predictive performance increases with increasing antibody diversity (Figure 4.2A).

[4]Changing the peptide library for each simulation run, but leaving assigned AAWS constant, does not significantly change Figure 4.2C.

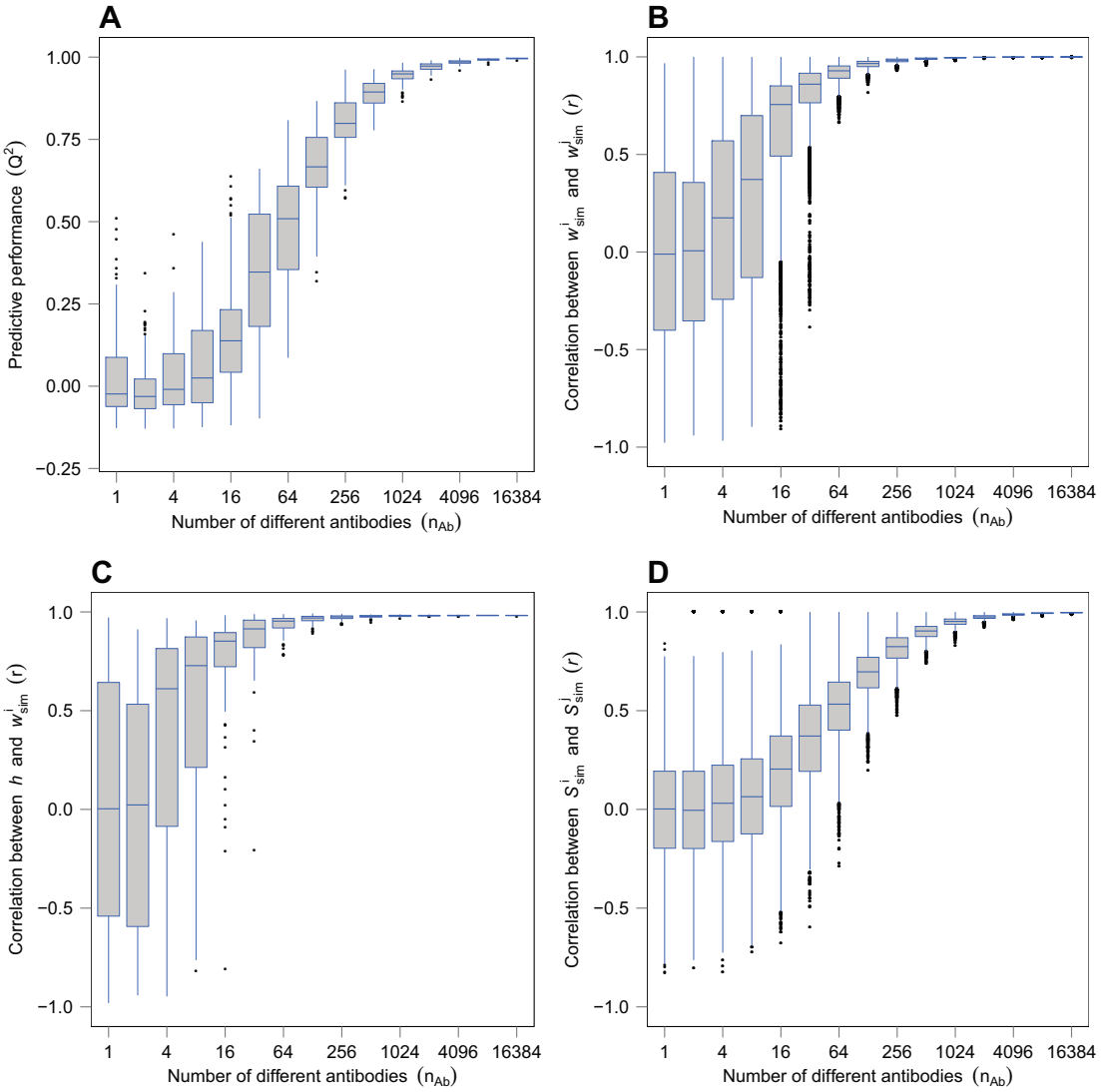


Figure 4.2: Simulations show that the predictive performance of antibody binding profiles improves with increasing antibody diversity. Antibody binding profiles ($\vec{S}_{\text{sim}}$) were simulated for antibody mixtures of 1 to 16384 different antibodies. (A) The predictive performance increases with increasing number of antibody variants (n_{Ab}), (B) as does the correlation (r) between all pairs of estimated AAWS ($\vec{w}^i_{\text{sim}}$), (C) between estimated AAWS and assigned AAWS ($\vec{h}$) and (D) between all pairs of the corresponding normalized signal intensity profiles ($\vec{S}^i_{\text{sim}}$). In (A–D), a random peptide library with associated AACM ($\mathbf{X}_{\text{sim}}$) of 255 14-mers and assigned AAWS ($\vec{h}$) were generated once and were kept constant across all simulation runs. For every mixture of n_{Ab}-different antibodies, 100 simulations with newly generated random antibody mixtures were run. Antibody binding profiles were computed using Equation 4.1. Corresponding AAWS ($\vec{w}^i_{\text{sim}}$) were determined using Equation 4.8.

4.3.4 Antibody dominance decreases the linear predictability of simulated signal intensity profiles

In the following, the effect of the selective increase in concentration of a subset of an antibody mixture (hereafter termed *dominant antibodies*) on predictive performance is studied. Based on the above results (Figure 4.2), one would expect that such a relative increase in antibody concentration has a decreasing effect on predictive performance. Antibody dominance would turn an unbiased mixture into a biased one.

In fact, simulations indicate that the reduction of predictive performance (Figure 4.3A) as well as the distancing of the signal intensity profiles from that of unbiased mixtures (Figure 4.3F) only occurs for those biased mixtures in which a small number of dominant antibodies is found in *absolute* terms (5–10 dominant antibodies). Indeed, for 1000 dominant antibodies in a biased mixture of 10000 antibodies, the predictive performance nears 1 (Figure 4.3A and D).

Of note, using Spearman correlation coefficients yields analogous results to those shown in Figure 4.3F: dominant antibodies induce *rank* changes which distance signal intensity profiles of biased from those of unbiased mixtures. Ranks of signal intensity profiles are determined by assigning the rank 1 to the peptide with the highest signal intensity, the rank 2 to the peptide with the second highest signal intensity and so forth.

Therefore, the fruitfulness of the term "antibody-dominated mixture"—implying a less than perfect predictive performance and lower recovery of assigned AAWS $\vec{h}$ (Figure 4.3C)—depends as much on the number of dominant antibodies as on their actual increase in concentration. Antibody dominance is, therefore, absent in a highly diverse randomly generated antibody mixture if for all different antibody concentrations $[\mathrm{Ab}]_k \in \mathrm{C}_j$ the diversity of antibodies tends to infinity ($n_{\mathrm{Ab}_{[\mathrm{Ab}]_k}} \to \infty$). This negative definition of antibody dominance is consistent with property (iii) of unbiased mixtures (Section 4.3.3, page 53).

Furthermore, Figure 4.3D[5] shows that simulated monoclonal antibodies exist which show quite high predictive performance values with almost perfect recovery of assigned AAWS (Figure 4.3E). These simulated monoclonal antibodies are studied in Section 7.2.

4.3.5 The signal of unbiased mixtures enables the isolation of the signal of simulated dominant antibodies in biased mixtures

The link between unbiased and biased mixtures are the dominant antibodies. One could therefore ask whether, given the signal of an unbiased mixture, the signal of the biasing dominant antibodies can be isolated from any biased mixture's signal.

Let $\vec{S}_{\mathrm{U}}$ be the signal intensity profile of an unbiased mixture AM_k of $n_{\mathrm{Ab,U}}$ antibodies. Let $\vec{S}_{\mathrm{U\text{-}D}}$ be the signal intensity profile of a biased mixture AM_j of $n_{\mathrm{Ab,U\text{-}D}}$ antibodies dominated by $n_{\mathrm{Ab,D}}$ antibodies. The antibody repertoires of AM_k and AM_j are not required to match ($\mathrm{A}_k \neq \mathrm{A}_j$). Let $\vec{S}_{\mathrm{D}}$ be the signal intensity profile of the $n_{\mathrm{Ab,D}}$ dominant

[5] Slight variations of predictive performance values by concentration (Figure 4.3D) are due to the R implementation of PLSR.

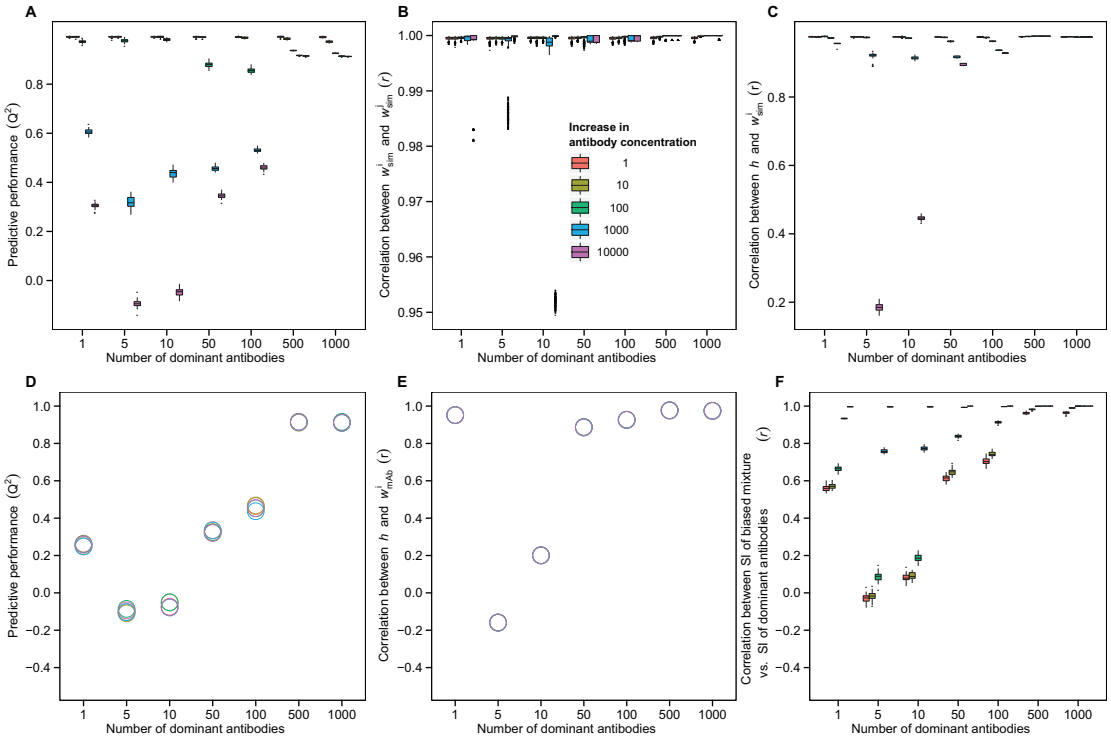


Figure 4.3: Simulations show that the predictive performance of biased mixtures changes in function of both the number and concentration of dominant antibodies. Simulations were done with a total of 10000 antibodies. Subsets of cardinalities 1–1000 were increased in concentration within the range of 1 (no increase) to 10000 (10000-fold increase of dominant antibodies with respect to the rest). (A) For all abundances of dominant antibodies, predictive performance (Q^2) is decreased with increasing antibody concentration. However, the degree to which predictive performance (Q^2) is decreased, depends on the number of dominant antibodies increased in concentration. (B) The correlation (r) between all pairs of estimated AAWS ($\vec{w}^i_{\text{sim}}$) is shown as is (C) the recovery of assigned AAWS ($\vec{h}$) by estimated AAWS ($\vec{w}_{\text{sim}}$). (D) Predictive performance (Q^2) values as well as (E) the recovery of assigned AAWS of dominant antibodies used to bias mixtures of Figures (A–C) are shown. (F) The Pearson correlation of normalized signal intensities obtained with dominant antibodies alone (D–E) and normalized signal intensities obtained of dominant antibodies in a biased mixture (A–C) is shown. Using Spearman correlation yields equivalent results. In (A–F), for every simulation run (50 per boxplot), the peptide library of 255 14-mers with the associated AACM ($\mathbf{X}_{\text{sim}}$) and the assigned AAWS $\vec{h}$ were kept constant. Simulated antibodies, but not the dominant ones, were changed for each simulation run. Antibody binding profiles were computed using Equation 4.1. Corresponding AAWS ($\vec{w}^i_{\text{sim}}$) were determined using Equation 4.8.

antibodies. $\vec{S}_{\text{U}}$, $\vec{S}_{\text{U-D}}$ and $\vec{S}_{\text{D}}$ are assumed to be simulated with the same peptide library. Also, it is assumed that $n_{\text{Ab,U}}, n_{\text{Ab,U-D}} \gg n_{\text{Ab,D}}$.

According to Equation 4.1, $S_{i,\text{U}} = \dfrac{s_{i,\text{U}}}{1 + s_{i,\text{U}}}$, and $s_{\text{U}} = -\dfrac{S_{i,\text{U}}}{S_{i,\text{U}} - 1}$, $S_{i,\text{U}} \neq 1$.

$S_{i,\mathrm{D}} = \frac{s_{i,\mathrm{D}}}{1+s_{i,\mathrm{D}}}$ and $s_{i,\mathrm{D}} = -\frac{S_{i,\mathrm{D}}}{S_{i,\mathrm{D}}-1}$, $S_{i,\mathrm{D}} \neq 1$.

Then, $S_{i,\mathrm{D}}$ can be expressed as follows:

$$S_{i,\mathrm{D}} \approx \frac{\frac{S_{i,\text{U-D}}-S_{i,\mathrm{U}}}{(S_{i,\text{U-D}}-1)(S_{i,\mathrm{U}}-1)}}{1+\frac{S_{i,\text{U-D}}-S_{i,\mathrm{U}}}{(S_{i,\text{U-D}}-1)(S_{i,\mathrm{U}}-1)}} \tag{4.9}$$

For an extended version of the derivation, please refer to the Appendix (Section A.8.3)

Applying Equation 4.9 to the simulations performed for Figure 4.3 shows that the quality of the isolation of the signal of the dominant antibodies increases with increasing concentration of dominant antibodies but worsens with an increasing number of dominant antibodies (Figure 4.4)[6].

Thus, as predicted by theory, a negative correlation between the predictive performance of a biased mixture (Figure 4.3A) and the isolation quality of the signal of the dominant antibodies that are in that biased mixture (Figure 4.4) exists. The higher the predictive performance of a biased mixture, the lower is the quality of signal isolation since an increasing predictive performance signifies a decreasing antibody dominance (Section 4.3.4).

In cases of lower concentrations of dominant antibodies, the isolation quality depends on the antibody composition of the given biased mixture as evidenced by the high variation[7] in correlation coefficients (Figure 4.4).

4.4 Summary

- A mathematical model of antibody-peptide binding was formulated (Section 4.2). Therein, a special case of simulated antibody mixtures, termed *unbiased* (Section 4.3.3), was found. Unbiased mixtures yield antibody binding profiles which are completely determined by assigned AAWS $\vec{h}$: the unbiased mixture's antibody composition becomes unimportant for peptide signal intensity (Figure 4.2D) as antibody diversity tends to infinity. This was shown both by mathematical analysis (Section 4.3, Equation 4.7) and simulations (Section 4.3.2, Figure 4.2).

- A regression model linking amino acid composition and signal intensity profiles was formulated (Section 4.3.2). It yields estimated AAWS ($\vec{w}$), which display the importance of every amino acid for signal intensity generation. For unbiased mixtures, the regression model's residuals tend to zero (Figure 4.2A). In this case, estimated AAWS are near perfect predictors of antibody binding profiles ($Q^2 \to 1$) and recover assigned AAWS ($\vec{h}$) (Figure 4.2C).

- The fruitfulness of the term "antibody-dominated mixture"—implying both a less than perfect predictive performance (Q^2) as well as recovery of assigned AAWS caused by a concentration increase of several antibodies in the mixture—depends

[6] If $S_{i,\text{U-D}} = S_{i,\mathrm{U}}$, then $S_{i,\mathrm{D}} = \vec{0}, \forall i \in \{1,\ldots,\#P\}$.

[7] Some of the variation may also be due to numerical reasons.

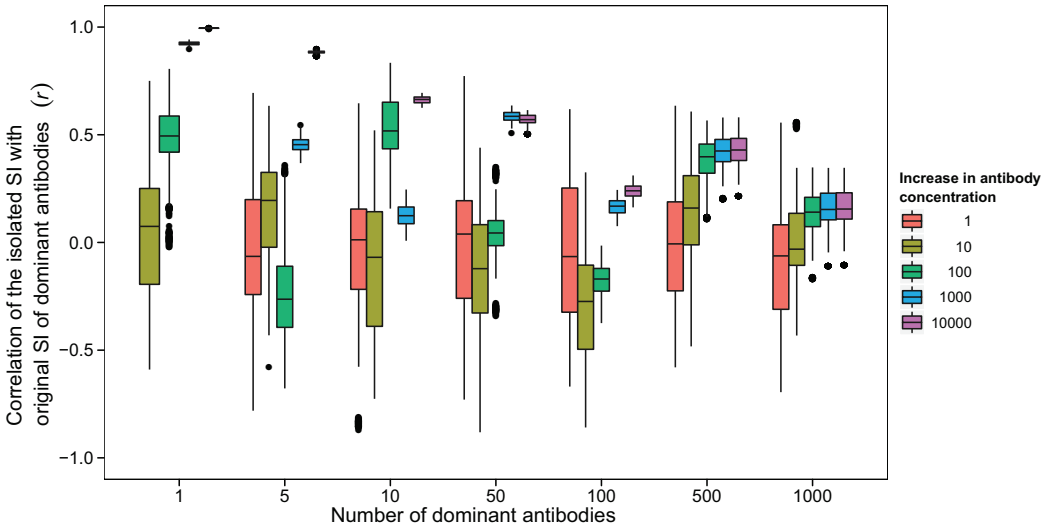


Figure 4.4: The quality of the isolation of the signal of the dominant antibodies ($\vec{S}_{\mathrm{D}}$) from the antibody binding profile of a biased mixture ($\vec{S}_{\mathrm{U\text{-}D}}$) increases with an increasing *concentration* of dominant antibodies but decreases with an increasing *number* of dominant antibodies. For lower antibody concentrations, the isolation quality depends on the unbiased mixture ($\vec{S}_{\mathrm{U}}$) as evidenced by the high variation in correlation coefficients. Pearson correlation coefficients between the signal of dominant antibodies isolated from the profile of the biased mixture and the simulated (original) signal intensities of the dominant antibodies are shown. Antibody binding profiles were simulated as in Figure 4.3 except that a normalization of profiles was *not* performed. Signal intensity profiles of dominant antibodies were isolated using Equation 4.9.

as much on the number of dominant antibodies as on their actual increase in concentration (Section 4.3.4, Figure 4.3A–C). Dominant antibodies distance profiles of biased antibody mixtures from those of unbiased ones by induction of rank changes (Figure 4.3F).

- Given the signal of the unbiased mixture, mathematical analysis indicates that the signal of the dominant antibodies can be, to a certain extent, isolated from the signal of any biased mixture (Section 4.3.5). Complementary simulations show that the quality of the isolation is dependent on both the number and the concentration of dominant antibodies: it is essentially negatively proportional to the biased mixture's predictive performance (Figures 4.3 and 4.4).

- Further mathematical derivations in the Appendix (Section A.8) show that (i) the simulation of signal intensities is not bijective. Input data such as antibody diversity (n_{Ab}) cannot be restored from the signal intensity (output) space (Section A.8.1, Footnote 22 page 131). (ii) The signal intensity subspace of unbiased mixtures is considerably reduced compared to that of biased mixtures (Section A.8.2, Equation S.7). This reduction depends both on peptide sequence length and the cardinality of the amino acid alphabet (#AA).

5 A minimal model of antibody-peptide binding: in vitro validation of mathematical predictions

Parts of this Chapter were recently published [104].

5.1 The predictive performance differs between monoclonal and serum-antibody binding profiles

In order to test in vitro the in silico prediction that the regression model's (Equation 4.8) (i) predictive performance, the pairwise correlation of both (ii) signal intensity profiles and (iii) AAWS increase with increasing antibody diversity (Figure 4.2), the predictive performance of 58 BALB/c mouse serum samples (antibody diversity $n_{\mathrm{Ab}} \gg 1$, Mouse study, Section 3.5.8) was compared with that of 13 human monoclonal IgG antibodies (antibody diversity $n_{\mathrm{Ab}} = 1$, Section 3.5.9). Both, monoclonal and serum antibodies were incubated on microarrays with the peptide library $J^{255}_{14\text{-mer}}$ (Table 3.1)[1]. Indeed, serum antibodies showed both a significantly higher predictive performance (median $Q^2 = 0.39$) (Figure 5.1A, $p < 0.001$) and significantly higher pairwise correlations between AAWS (Figure 5.1B, $p < 0.001$) than monoclonal antibodies (median $Q^2 = 0$), which confirms the predictions of the mathematical model (Section 4.3, Figure 4.2). The median pairwise Pearson correlation[2] of normalized monoclonal antibody binding profiles was found to be $r = 0.40$, whereas that of BALB/c sera was determined to be $r = 0.75$ (Figure 5.4).

5.2 Predictive performance decreases in the course of an HB-infection

In order to quantify the influence of immune response stage during the infection with the murine parasite *Heligmosomoides bakeri* (HB, Nematoda) on predictive performance, the mouse serum samples were divided into three groups: *healthy*, (*HE*), *acute phase* (*AP*, 10 and 14 dpi) and *early chronic phase* (*CP*, 18 dpi) (Section 3.5.8, Figure 3.1) [287]. In the course of the immune response, the predictive performance (Figure 5.2A) and

[1] Predictive performance values before and after secondary-antibody correction of both monoclonal and serum incubations can be found in Figure S.4.

[2] The impact of the secondary detection antibody on *monoclonal* signal intensity was shown to be large (Figure S.4). This is in contrast to the secondary detection antibody's impact on signal intensity profiles of *serum* antibodies (Section 3.5, page 32, Figure S.4).

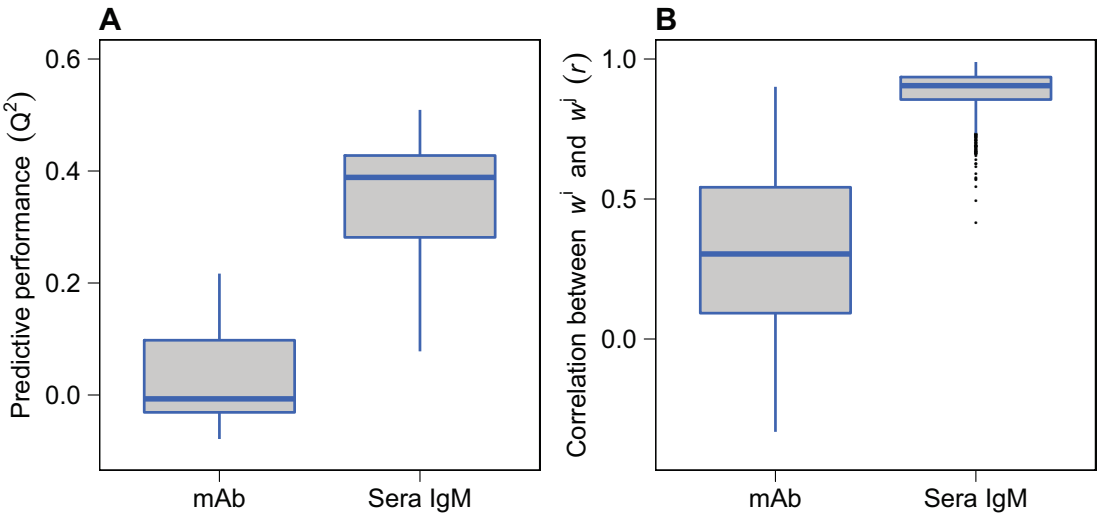


Figure 5.1: Predictive performance values and pairwise correlation of AAWS are higher for serum IgM than for monoclonal antibodies. (A) Predictive performance values were calculated for human monoclonal IgG (mAb, Section 3.5.9) and serum IgM antibody (Sera IgM) binding profiles (BALB/c mice, Mouse study, Section 3.5.8). (B) The pairwise Pearson correlation (r) of the corresponding AAWS $\vec{w}^j$ is shown. Sample numbers for the respective groups are: mAb, 13; sera IgM, 58. Differences between monoclonal and serum IgM antibodies in predictive performance (Q^2) and pairwise correlation (r) of AAWS are significant ($p < 0.001$). Antibody binding profiles were measured with the $J^{255}_{\text{14-mer}}$ library. Corresponding AAWS ($\vec{w}^j$) were determined using Equation 4.8.

pairwise correlation of AAWS decrease significantly (Figure 5.2B) as does the pairwise correlation of non-normalized IgM signal intensity profiles (Table 5.1). The decrease in pairwise correlation of signal intensity profiles is due to biological (Table 5.1) and not technological variability (Table 5.2): the intra-group median correlation of *HE* IgM signal intensity profiles ($r = 0.83$; $P < 0.001$) is inferior to intra- and inter-array correlation of *HE* sera (Tables 5.2 and 5.1)

5.3 Stages of murine immune response differ in their amino acid-associated weights and signal intensity profiles

In order to test whether the AAWS determined for all 58 BALB/c mouse serum samples differ systematically by stage of immune response (*HE*, *AP*, *CP*), principal component analysis (PCA, Section 3.8.1) was used. Together, the first two principal components yield a strong separation of healthy (*HE*) and diseased mice (*AP*, *CP*). Also, *AP* and *CP* samples separate (Figure 5.3). Thus, during an immune response against HB, AAWS change in a systematic way.

Likewise, (i) the corresponding IgM signal intensity profiles of sera of the different stages of immune response as well as (ii) their ranks can be separated by PCA (Figures S.2 and S.3).

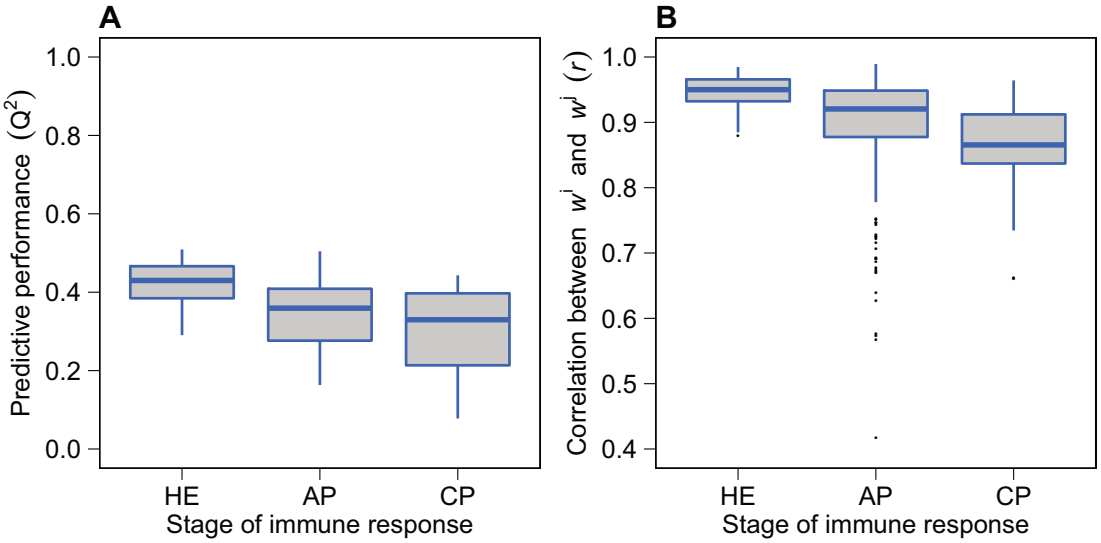


Figure 5.2: Predictive performance and pairwise correlation of AAWS decrease for serum IgM antibodies in the course of the HB-infection. (A) Predictive performance values (Q^2) were computed from serum IgM antibody binding profiles across three stages of immune response: *healthy* (*HE*), *acute* (*AP*), *early chronic* (*CP*). (B) The pairwise correlation (r) of the corresponding AAWS $\vec{w}^j$ is shown. Numbers of BALB/c mouse serum samples: 15 samples from *HE* mice; after infection with HB: 15 samples at 10 dpi and 13 samples at 14 dpi (*AP*), and 15 samples at 18 dpi (*CP*) totaling 58 BALB/c mouse serum samples. Differences in predictive performance (Q^2) between *HE* and both *AP* and *CP* mice are significant ($p < 0.01$), as are differences in pairwise correlation (r) between all three stages of immune response ($p < 0.001$). Antibody binding profiles were measured with the $J^{255}_{14\text{-mer}}$ library. Corresponding AAWS ($\vec{w}^j$) were computed using Equation 4.8.

In order to characterize stages of immune response further, nested leave-one-out cross-validation (LOOCV) using P-SVM was employed [278] (Section 3.8.2). P-SVM was applied to the IgM signal intensity profiles of the three subproblems (*HE–AP*, *HE–CP*, *AP–CP*). Balanced accuracies (BACC, Section 3.8.2, Equation 3.7) range from 80.0% to 100% (Table 5.3). Notably, in order to obtain compact models only a small set of peptides was used and all parameter combinations in the inner cross-validation loop for which more than three models exceeded an upper limit of six selected peptides were rejected (Section 3.8.2). The validation of the significance of the P-SVM classification results was performed with permutation testing (Section 3.8.2). All three subproblems led to $p < 0.05$ (Table 5.3).

Additionally, PSVM-nested cross-validation was applied to the ranks of IgM profiles. BACCs are comparable to those found for signal intensities (83.3%–96.5%, Table 5.3). For the subproblem *HE–CP*, ranks perform better than signal intensities (Table 5.3). All three subproblems yield $p < 0.01$ (Table 5.3).

The successful classification of subproblems shows that the changes in binding patterns induced by the immune response against HB are *consistent* across individuals of a group (*HE*, *AP*, *CP*).

	Median intra-group correlation coefficient (r)	Median inter-group correlation coefficient (r)
HE	0.83	
AP	0.79	
CP	0.69	
HE–AP		0.75
HE–CP		0.71
AP–CP		0.72

Table 5.1: Biological variability of random-sequence peptide array probing increases with progression of HB-infection. To evaluate changes in serum IgM profiles during an HB-infection (BALB/c mice, Mouse study, Section 3.5.8), the biological variability of non-normalized IgM signal intensity profiles within and among three groups of sera was investigated: pre-infection sera (*HE*), acute infection sera (*AP*) and early chronic infection sera (*CP*) (Figure 3.1A). Median intra-group correlation is highest for *HE*. Progression of infection is associated with a loss of correlation within the groups of *AP* and *CP*. Median intra-group correlation was determined by the median of pairwise non-redundant correlation coefficients between all serum samples of one group ($p < 0.001$ for all individual Pearson correlation coefficients). Inter-group correlation was calculated by taking the median of the correlation coefficients between sub-arrays of microarray pairs ($p < 0.001$ for all individual Pearson correlation coefficients). Differences between intra-group Pearson correlation coefficients are significant ($p < 0.001$). Sample numbers for the respective groups are: *HE*, 15; *AP*, 28; *CP*, 15.

	Median intra-array correlation coefficient (r)	Median inter-array correlation coefficient (r)
901-18	0.96	
901-19	0.94	
901-20	0.93	
901-18–901-19		0.90[a]
901-18–901-20		0.91
901-19–901-20		0.88

[a]Technological variability was also assessed for all other experimental studies with repeated measurements of healthy control sera. It was found to be generally above $r_{\text{Pearson}} = 0.85$ (Section 3.5).

Table 5.2: Assessment of the technological variability of random-sequence peptide array probing with the $J^{255}_{14\text{-mer}}$ library. To establish the technological variability of the experimental setup of the Mouse study (Section 3.5.8), 3×5 sub-array panels with the same serum obtained from an uninfected (*HE*) BALB/c mouse were probed to evaluate inter- and intra-array correlation of non-normalized IgM signal intensity profiles. For intra-array correlation, pairwise non-redundant correlation coefficients between the five sub-array panels of each array were calculated and the median was determined (Pearson $r > 0.92$; $p < 0.001$). Inter-array correlation, calculated by taking the median of the correlation coefficients between sub-arrays of microarray pairs, was found to be lower than intra-array correlation (Pearson $r > 0.87; p < 0.001$). Row names denote the array identifiers.

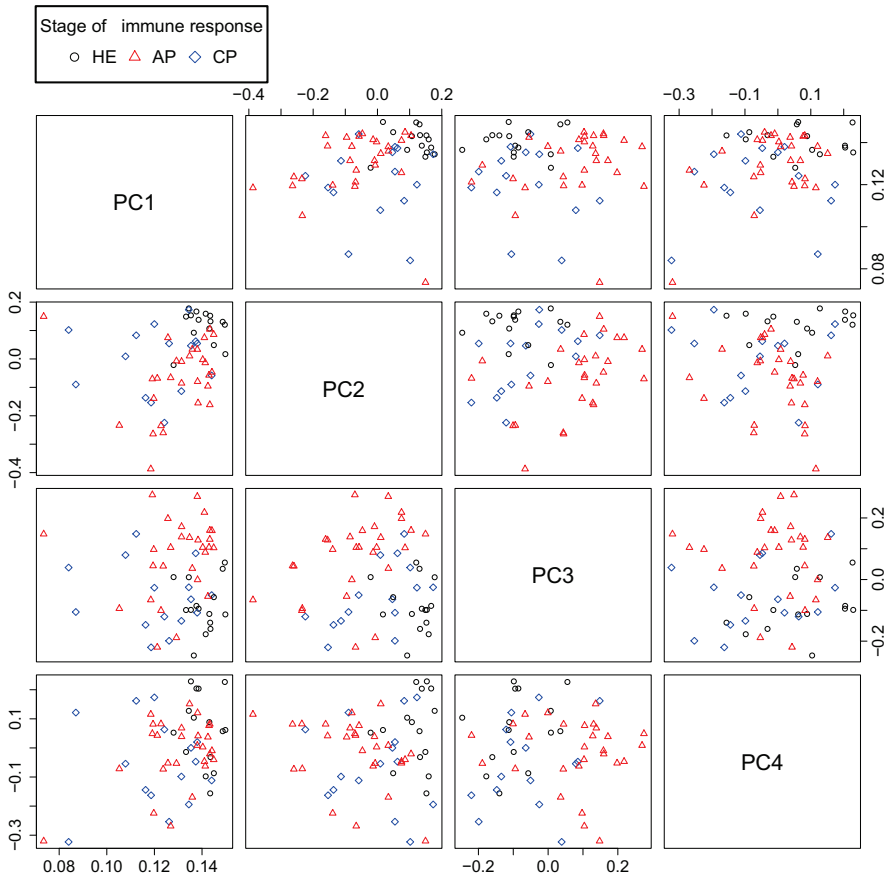


Figure 5.3: Stages of immune response differ in their AAWS. Principal component analysis was applied to the 255×58 signal intensity matrix (255 analyzed peptide signal intensities of library $J^{255}_{14\text{-mer}}$ times 58 BALB/c samples of the Mouse study, Section 3.5.8). The loadings of the first 4 principal components are shown. The first two principal components (PC1, PC2) separate *HE* and diseased (*AP*, *CP*) mice whereas the second and the third principal component (PC2, PC3) tend to separate *AP* and *CP* samples. Four components (PC1–PC4) explain 96.2% of the variance in the data. Sample numbers for the respective groups are: *HE*, 15; *AP*, 28; *CP*, 15.

This significant consistency may not be with respect to the entire IgM profile or ranks but certainly with regard to the small percentage of the library used by P-SVM. In fact, removing the highly discriminatory peptides by subproblem from the data set (Section 3.8.2), considerably reduces the BACC of the subproblems *HE–CP* and *AP–CP* both for IgM signal intensity profiles and their ranks (Table S.1). Only the selected peptides for the subproblem *HE–AP* were not unique in their ability to classify IgM profiles and ranks by stage of immune response with high accuracy (Table S.1).

IgM signal intensity profiles				
Subproblem	BACC [%]	Sensitivity [%]	Specificity [%]	Significance (p-value)
HE–AP	100	100	100	0
HE–CP	80.0	66.7	93.3	0.04
AP–CP	84.4	86.7	82.1	0
Ranks				
Subproblem	BACC [%]	Sensitivity [%]	Specificity [%]	Significance (p-value)
HE–AP	96.5	92.9	100	0
HE–CP	83.3	73.3	93.3	0.006
AP–CP	89.8	86.6	92.9	0

Table 5.3: Assessment of the P-SVM balanced classification accuracy (BACC) for both IgM signal intensity profiles and their ranks of subproblems of the Mouse study (BALB/c, Section 3.5.8, Figure 3.1). Signal intensity profiles were determined with the $J^{255}_{\text{14-mer}}$ library. All BACCs are higher than 83%. Determined BACCs were found to be significant ($p < 0.05$). Sample numbers for the respective groups are: *HE*, 15; *AP*, 28; *CP*, 15.

5.4 Assessment of predictive performance values and pairwise correlation of estimated AAWS and signal intensity profiles by experimental study

In order to broaden the perspective beyond the Mouse study described in the previous Sections, predictive performance values (Q^2, Figure 5.4A), estimated AAWS ($\vec{w}$, Figure 5.4B) and signal intensity profiles of healthy individuals[3] ($\vec{S}$, Figure 5.4C) were assessed for all experimental studies described in *Methods* (Section 3.5): median predictive performance values, analyzed by study, range from $Q^2 = 0.01$ (NephroFIT study, human, 15-mers) to $Q^2 = 0.46$ (NOD study, NOD mice), median pairwise Pearson correlation of AAWS ranges from $r = 0.63$ (NephroFIT study, human, 13-mers) to $r = 0.99$ (NOD study, NOD mice, 13-mers) and median pairwise Pearson correlation of signal intensity profiles from $r = 0.48$ (NephroFIT study, human, 13-mers) to $r = 0.95$ (NOD study, C57BL/6, 13-mers) (Figure 5.4)[4].

Taken together, (i) murine samples show higher predictive performance values than human samples (also on 15-mer libraries), (ii) on JPT arrays 13-mers yield consistently higher Q^2-values than 15-mers, (iii) and with respect to 15-mers, Pepscan arrays show higher Q^2-values than JPT arrays. (iv) However, independent of the manufacturer (JPT, Pepscan), pairwise Pearson correlation of AAWS by study is relatively high across all

[3] For both the NephroFIT and the NephroFIT-Pepscan study, "No rejection" individuals were assessed in this Figure, because no more than two samples of healthy controls were incubated in these studies. "No rejection"-AAWS are highly correlated to AAWS of healthy controls (data not shown).

[4] Ranges of correlation coefficients with respect to both AAWS (Figure 5.4B) and signal intensity profiles (Figure 5.4C) were found to be similar for Spearman correlation.

studies with murine samples showing highest correlation. The same is true[5] for signal intensity profiles (Figure 5.4).

Please refer to Chapter 8 for a comparative analysis of AAWS *across* experimental studies.

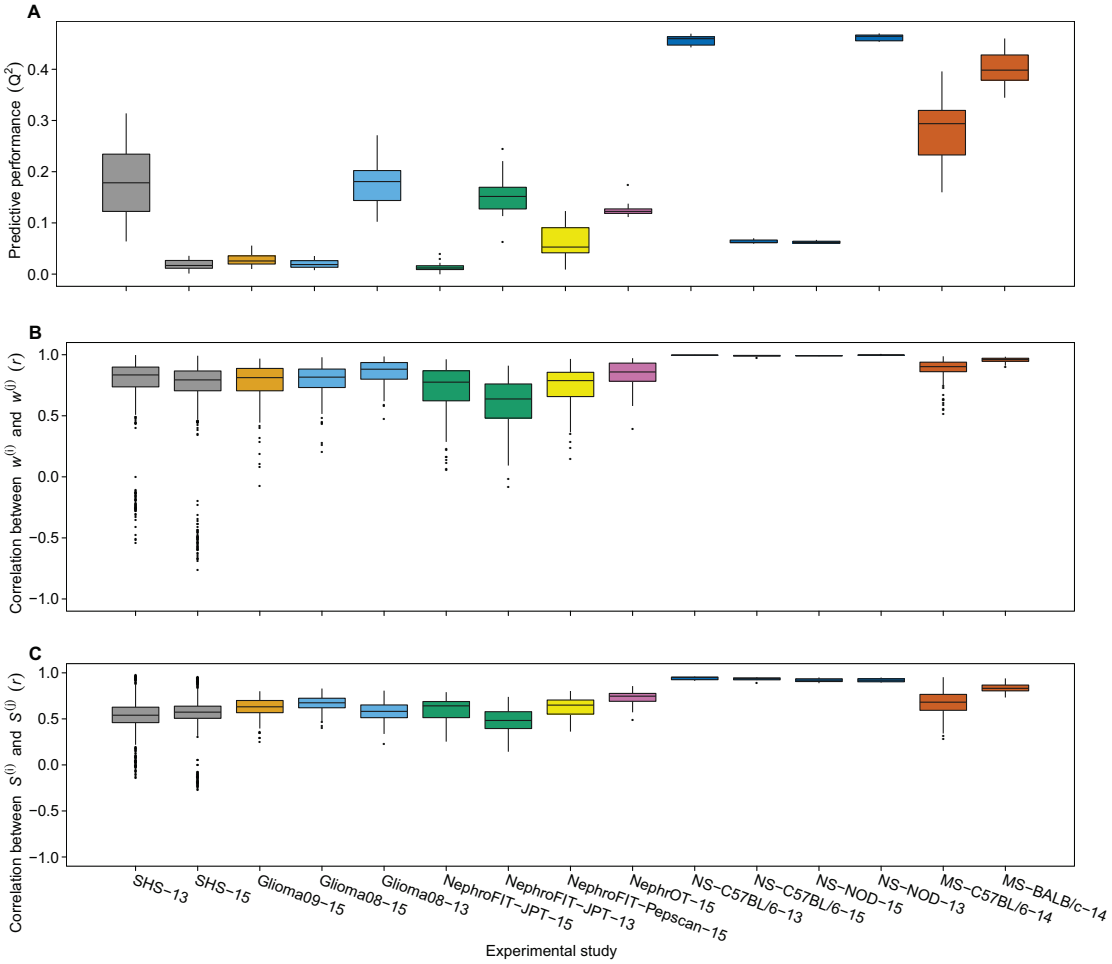


Figure 5.4: Assessment of (A) predictive performance values (Q^2), (B) pairwise Pearson-correlation (r) of estimated AAWS ($\vec{w}$), as well as of (C) normalized IgM signal intensity profiles ($\vec{S}$) of *healthy* individuals by experimental study (Section 3.5). (For both the NephroFIT and the NephroFIT-Pepscan study, "No rejection" individuals were assessed in this Figure, because no more than two samples of healthy controls were incubated in these studies). Generally, the median Q^2 is highest for murine samples and lower peptide lengths (13-mer versus 15-mer). Pepscan arrays show highest Q^2-values for arrays with 15-mers. AAWS of each sample were determined with Equation 4.8. Differences in predictive performance values between 13- and 15-mers are significant ($p < 0.05$). Average AAWS are named according to the convention: Experimental study-(Array manufacturer/Mouse model)-Peptide length.

[5]The minimum median correlation of $r = 0.48$ of serum signal intensity profiles is higher (i) than that of monoclonal antibodies ($r = 0.40$, Section 5.1) and (ii) blank–serum correlation (Section 3.5).

5.5 Assessment of the correlation of average AAWS with both propensity scales for epitope prediction and amino acid physico-chemical properties

Because of the high pairwise correlation of healthy AAWS found throughout all experimental studies[6], they were averaged by experimental study (Figure 5.5). Average AAWS differ mostly by species (human, mouse) and array platform (JPT, Pepscan) (Figure 5.5). Whereas for murine samples, aromatic amino acids as well as methionine (M) and glutamic acid (E) are top positive contributors to signal intensity, top ranking amino acids for human samples incubated on JPT arrays are methionine (M) and proline (P). Average AAWS of human samples incubated on Pepscan arrays showed mainly lysine (K), histidine (H), proline (P) and tryptophane (W) as positively contributing amino acids. For a further characterization of the technological variability of AAWS, consult Chapter 8.

AAWS represent a priority scale for peptide-antibody binding assigning to every amino acid the importance of contribution to the measured (or simulated) signal intensity. In addition, analogously to QSAR modeling, AAWS can a posteriori be conceived of as a vector representing correlates of the respective amino acids' physico-chemical (PC) properties. Therefore, average AAWS (Figure 5.5) were Spearman correlated with the z-scale developed by Sandberg and colleagues (Table 5.4, [191]). The z-scale aggregates in matrix form 26 PC amino acid properties for each of the 20 examined amino acids (Table 3.2). The NephrOT study shows the highest number of correlation coefficients exceeding a Spearman correlation of 0.5 (Table 5.4).

In order to compare the average AAWS with other published amino acid scales for epitope prediction, they were Spearman correlated with four propensity scales (Section 1.7.3) published by Parker and colleagues [194] (hydrophilicity), Kolaskar and Tongaonkar [288] (antigenicity), Chou and Fasman [289] (secondary structure) and by Emini and colleagues [197] (accessibility). The correlation of average AAWS with the described propensity scales was found to be generally poor except for AAWS originating from Pepscan arrays (NephroFIT-Pepscan, NephrOT, Table 5.5). Correlating the four mentioned propensity scales among each other leads to coefficients ranging between $r_{\text{Spearman}} = -0.56$ and $r_{\text{Spearman}} = 0.68$.

5.6 Summary

- The regression model (Equation 4.8) shows higher median predictive performance values for serum than for monoclonal antibodies (Section 5.1, Figure 5.1). Furthermore, the pairwise correlation of serum signal intensity profiles and AAWS is significantly higher than that of monoclonal antibodies (Section 5). Thus, in vitro data are in agreement with the predictions of the mathematical model (Figure 4.2).

[6] Except for the Mouse study, there are mostly no significant differences between 15-mer-AAWS between different diagnosis groups of human studies (data not shown). This may be primarily explained be the low a priori predictive performance (Q^2) of 15-mers (Figure 5.4A).

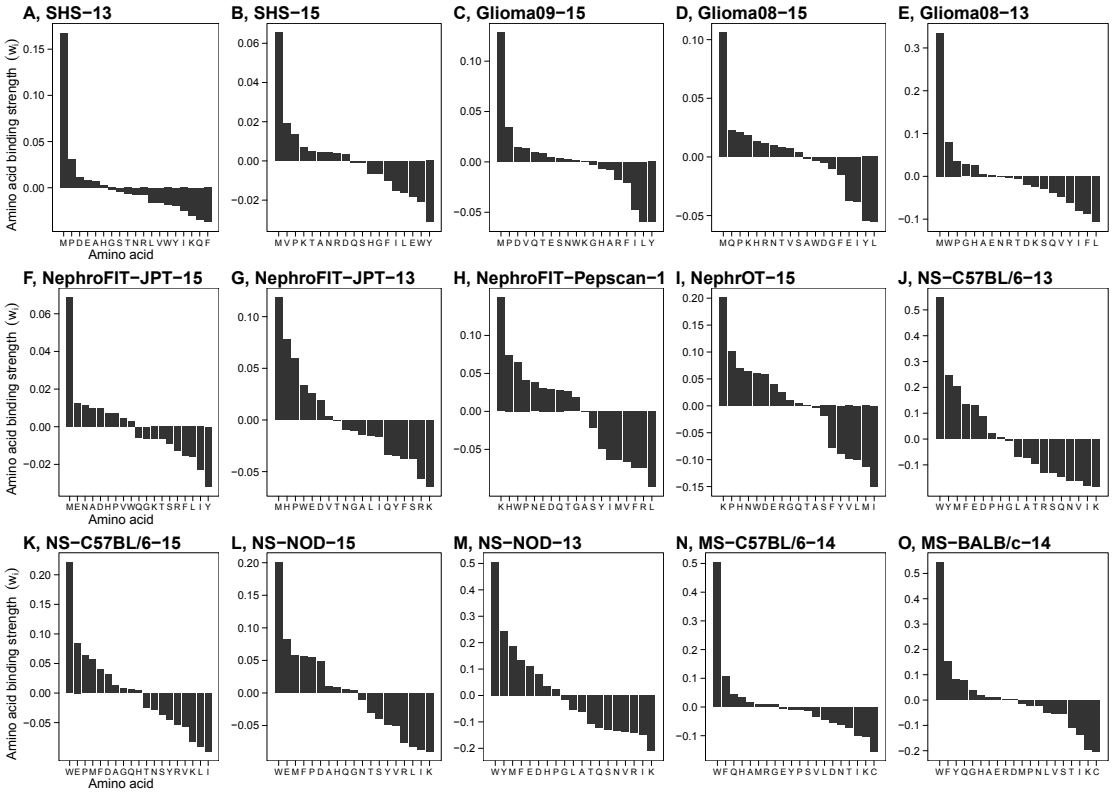


Figure 5.5: Average AAWS of healthy individuals differ by species (human, mouse) and manufacturer (JPT, Pepscan). AAWS of *healthy* individuals were determined with Equation 4.8 and averaged by experimental study. (For both the NephroFIT and the NephroFIT-Pepscan study, "No rejection" individuals were assessed in this Figure because no more than two samples of healthy controls were incubated in these studies). The difference between two weights (w_i) indicates the contribution to the difference in normalized signal intensity corresponding to an amino acid substitution. Top ranking amino acids are: (A–G) (human samples, JPT arrays): M, P; (H–I) (human samples, Pepscan arrays): K, H, P, W; (J–0) (murine samples, JPT (contact and non-contact printed)): F, W Y, E, M. Average AAWS are named according to the convention: Experimental study-(Array manufacturer/Mouse model)-Peptide length.

- Median predictive performance decreases significantly in the course of a murine HB-infection (Section 5.2, Figure 5.2A) as does the pairwise correlation of both estimated AAWS (Figure 5.2B) and signal intensities (Table 5.1).

- Stages of immune response do not only differ by AAWS (Figure 5.3) but also by IgM signal intensity profiles (Figure S.2) and their ranks (Figure S.3): changes in binding patterns induced by the immune response against HB are consistent across mice.

- In addition, size-restricted sets of selected peptides showed uniformly high prediction accuracies both for IgM profiles (Section 5.3, Table 5.3) and ranks (Table 5.3) of sera of healthy and HB-infected mice. Removing the selected peptides from the data

Experimental study	Number of absolute Spearman correlation coefficients higher than 0.5	Physico-chemical properties
SHS-13	3	vdW, POLAR, Stot
SHS-15	0	
Glioma09-15	0	
Glioma08-15	3	TL1, TL2, TL5
Glioma08-13	3	TL1, TL2, TL5
NephroFIT-JPT-15	0	
NephroFIT-JPT-13	0	
NephroFIT-Pepscan-15	2	TL5, TL6
NephrOT-15	8	TL1, TL2, TL4, TL5, TL6, TL7, logP, HDONR
NS-C57BL/6-13	4	TL1, NM1, ELUMO, HA
NS-C57BL/6-15	2	NM1, ELUMO
NS-NOD-15	4	NM1, NM7, NM12, ELUMO
NS-NOD-13	5	TL1, NM1, NM12, logP, ELUMO
MS-C57BL/6-14	0	
MS-BALB/c-14	0	

Table 5.4: Assessment of the Spearman correlation of average AAWS with amino acid physico-chemical (PC) properties (z-scale, Sandberg *et al.* [191]). The z-scale aggregates in matrix form 26 physico-chemical amino acid properties (listed below) for each of the 20 (19) (Table 3.2) examined amino acids. AAWS of the *healthy* individuals (or "No rejection" for NephroFIT-studies) of each experimental study were determined with Equation 4.8, averaged and Spearman-correlated with the z-scale. Only those PC properties are shown, which showed absolute Spearman-correlation coefficients higher than 0.5. The maximum absolute Spearman correlation coefficient is $r = 0.74$ (NephrOT-15, TL5). Correlation coefficients above $r = 0.5$ are significant ($p < 0.05$). MW (molecular weight), TLx (thin layer chromatography at various conditions), vdW (side chain van der Waals volume), NMx (NMR-proton shift at pD = x), logP (10log (octanol/water) partition coefficient), EHOMO (energy of highest occupied molecular orbital), ELUMO (energy of lowest unoccupied molecular orbital), HOF (heat of formation), POLAR (α-polarizability), EN (absolute electronegativity), HA (absolute hardness), Stot (total accessible molecular surface area), Spol (polar accessible molecular surface area), Snp (non-polar accessible molecular surface area), HDONR (number of hydrogen bond donors), HACCR (number of hydrogen bond acceptors), Chpos (indicator of positive charge in side chain), Chneg (indicator of negative charge in side chain). Average AAWS are named according to the convention: Experimental study-(Array manufacturer/Mouse model)-Peptide length.

set, decreases classification accuracy (Table S.1) of two of the three investigated subproblems. Thus, P-SVM-selected peptides are partly unique in their ability to classify IgM profiles and ranks with high classification accuracy.

- Median predictive performance values are higher for murine than for human samples. AAWS from 13-mer libraries show higher Q^2-values than those of 15-mers. Pepscan arrays show highest Q^2-values for human samples with respect to 15-mers (Section 5.4, Figure 5.4A). The correlation among AAWS of a given study is high (Figure 5.4B) as is that of IgM signal intensity profiles (Figure 5.4C), although to a lesser extent. Thus, the mathematical predictions (Figure 4.2)—pertaining to a certain

Experimental study	Chou et al. (Secondary structure, [289])		Emini et al. (Accessibility, [197])		Kolaskar et al. (Antigenicity, [288])		Parker et al. (Hydrophilicity, [194])	
	r	p	r	p	r	p	r	p
SHS-13	0.22	0.37	0.06	0.80	-0.44	0.06	0.33	0.16
SHS-15	0.05	0.82	0.10	0.68	-0.20	0.41	0.19	0.43
Glioma09-15	0.26	0.29	0.21	0.38	-0.45	0.05	0.40	0.09
Glioma08-15	0.23	0.35	0.39	0.10	-0.32	0.19	0.23	0.35
Glioma08-13	0.33	0.17	0.18	0.45	-0.60	0.01	0.14	0.56
NephroFIT-JPT-15	0.14	0.56	0.20	0.42	-0.56	0.01	0.37	0.12
NephroFIT-JPT-13	-0.03	0.91	-0.18	0.46	-0.22	0.36	-0.08	0.74
NephroFIT-Pepscan-15	0.56	0.01	0.52	0.02	-0.33	0.17	0.43	0.07
NephrOT-15	0.64	0	0.69	0	-0.41	0.08	0.49	0.03
NS-C57BL/6-13	0.05	0.84	-0.06	0.80	-0.15	0.54	-0.29	0.23
NS-C57BL/6-15	0.22	0.37	0.11	0.65	-0.47	0.04	0.07	0.77
NS-NOD-15	0.10	0.68	0.02	0.94	-0.41	0.08	0.00	1.00
NS-NOD-13	0.05	0.83	-0.11	0.66	-0.11	0.67	-0.27	0.26
MS-C57BL/6-14	-0.19	0.41	0.08	0.75	-0.16	0.51	-0.23	0.33
MS-BALB/c-14	0.02	0.94	0.19	0.43	-0.18	0.45	-0.13	0.60

Table 5.5: Assessment of the correlation of average AAWS with selected amino acid propensity scales for epitope prediction. AAWS of the *healthy* individuals (or "No rejection" for NephroFIT-studies) of each experimental study were determined with Equation 4.8, averaged and Spearman correlated with the listed propensity scales. Pepscan-AAWS (NephroFIT-Pepscan and NephrOT) show the highest number of significant correlation coefficients ($p_{\text{Spearman}} < 0.05$). Average AAWS are named according to the convention: Experimental study-(Array manufacturer/Mouse model)-Peptide length. Propensity scales were obtained from `http://tools.immuneepitope.org/tools/bcell/`.

independence of antibody mixture and measured signal intensity profile—are valid for both *human* and *murine* serum antibodies.

- AAWS of healthy individuals were averaged by experimental study. Average AAWS differ by species and array platform (mouse vs. human, Pepscan vs. JPT, Section 5.4, Figure 5.5).
- Average AAWS are poorly correlated with amino acid physico-chemical properties. However, NephrOT-AAWS are the exception with 8 out of 26 absolute Spearman correlation coefficients higher than 0.5 (Section 5.5, Table 5.4).
- Average AAWS of most experimental studies show poor correlations with widely used propensity scales for epitope prediction. The exception are Pepscan-AAWS, which show absolute Spearman correlation coefficients higher than 0.5 and have the highest number of significant correlation coefficients (Section 5.5, Table 5.5).

6 A minimal model of antibody-peptide binding: analysis of the impact of model parameters on signal intensity profiles

In the following, the influence of model parameters on simulated signal intensity profiles and predictive performance values is assessed. If applicable, experimental data are used to determine whether changes of parameters entail collinear results in silico and in vitro.

6.1 Simulations show that the impact of both peptide length and library size on predictive performance and recovery of assigned AAWS is minimal

For unbiased mixtures, simulations show that neither the peptide library size (50–5000 randomly generated peptides) nor the peptide length (5–25 amino acids) greatly impact predictive performance values. Across all tested library sizes and peptide lengths, the predictive performance remains above $Q^2 = 0.95$ (Figure 6.1).

Nevertheless, increasing the peptide library size has a slightly positive effect on predictive performance whereas higher peptide lengths have a slightly decreasing effect on predictive performance (Figure 6.1).

6.2 Assessing the impact of total antibody concentration on signal intensity and predictive performance

In this Section 6.2, a constant *relative* concentration within antibody mixtures is assumed in order to study the impact of *total* antibody concentration on peptide signal intensity. Assuming, without loss of generality, (i) that for given assigned AAWS $\vec{h}$, the peptide string $\vec{p}^{\,i}$ is the null vector ($\vec{p}^{\,i} = \vec{0}$) and (ii) that a given antibody mixture AM_1 is composed of a single antibody $\vec{a}^{\,1}$ such that the total antibody concentration is equal to the concentration of antibody $\vec{a}^{\,1}$ ($[\mathrm{Ab}]_{\mathrm{Total}} = [\mathrm{Ab}]_1$). Then, the signal intensity S_i for peptide $\vec{p}^{\,i}$ computes as follows:

$$S_i = \frac{[\mathrm{Ab}]_1 * \exp(0)}{1 + [\mathrm{Ab}]_1 * \exp(0)} = \frac{[\mathrm{Ab}]_1}{1 + [\mathrm{Ab}]_1} = \frac{[\mathrm{Ab}]_{\mathrm{Total}}}{1 + [\mathrm{Ab}]_{\mathrm{Total}}}. \tag{6.1}$$

Thus, the simulated signal intensity is a monotone increasing nonlinear function of the total antibody concentration, with signal intensities either saturating ($S_i \to 1$) if the total

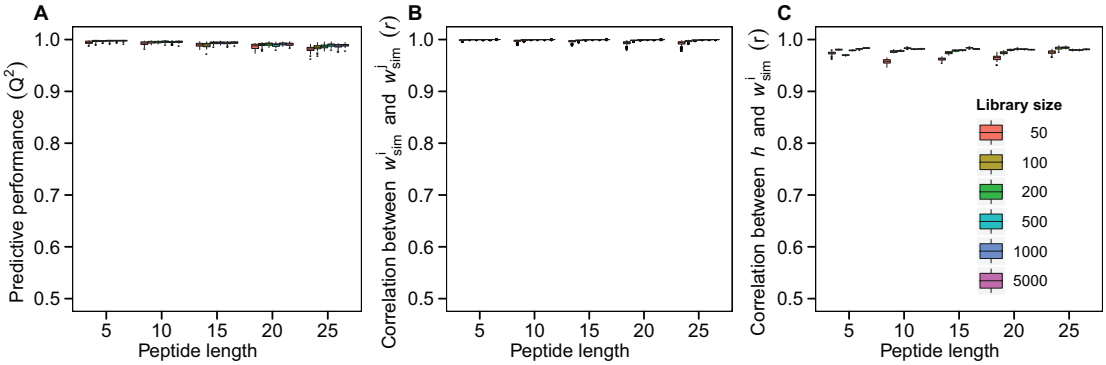


Figure 6.1: Simulations show that for unbiased mixtures (n_{Ab}=10000) the predictive performance is decreased for lower peptide library sizes and higher peptide length. However, for both tested parameters, predictive performance does not fall below $Q^2 = 0.95$. (A) Predictive performance (Q^2) decreases slightly with increasing peptide length but increases (slightly) with library size. (B) The same is true for the correlation (r) between all pairs of estimated AAWS $\vec{w}^i_{\text{sim}}$ as well as (C) for the recovery of assigned AAWS $\vec{h}$. In (A–C), random peptide libraries with associated AACMs ($\mathbf{X}_{\text{sim}}$) of 50–5000 peptides with 5–25-mers and assigned AAWS ($\vec{h}$) were simulated once and kept constant across all respective simulation runs. For each peptide library size and each peptide length, 50 simulations with newly generated unbiased mixtures were run. Antibody binding profiles were computed using Equation 4.1. Corresponding AAWS ($\vec{w}^i_{\text{sim}}$) were determined using Equation 4.8.

antibody concentration tends to infinity ($[\text{Ab}]_{\text{Total}} \to \infty$) or tending to zero ($S_i \to 0$) if the total antibody concentration tends to zero ($[\text{Ab}]_{\text{Total}} \to 0$) (Figures 6.2 and 6.3).

In fact, also a peptide library's mean simulated signal intensity ($\sum_{i=1}^{\#P} S_{\text{sim},i}$) is a monotone increasing function of the total antibody concentration (Figure 6.4A). Therefore, ranks of (mean) simulated signal intensities are not affected by a varying total antibody concentration (Figure 6.4B). Ranks are total antibody concentration ($[\text{Ab}]_{\text{Total}}$) invariant.

Correspondingly, a significant dependence of total antibody concentration on in vitro mean signal intensity ($\langle\vec{S}\rangle$) is found for all but one tested experimental study with Pearson correlation coefficients between signal intensity and total IgM concentration ranging from $r = 0.26$ (Mouse Study) to $r = 0.80$ (NephrOT study) (Figure 6.5 and Table 6.1).

Varying the total antibody concentration does not alter the predictive performance of simulated antibody binding profiles: across all tested values of $[\text{Ab}]_{\text{Total}}$ (0.01–10000), the predictive performance is high for unbiased mixtures (Figure 6.6B) and low for monoclonal antibodies (Figure 6.6A). The slight decrease of predictive performance values with growing total antibody concentration is due to the saturation of signal intensities (Figure 6.2).

Accordingly, in vitro measurements show no evidence for a significant dependence of predictive performance on total antibody concentration (Table 6.1).

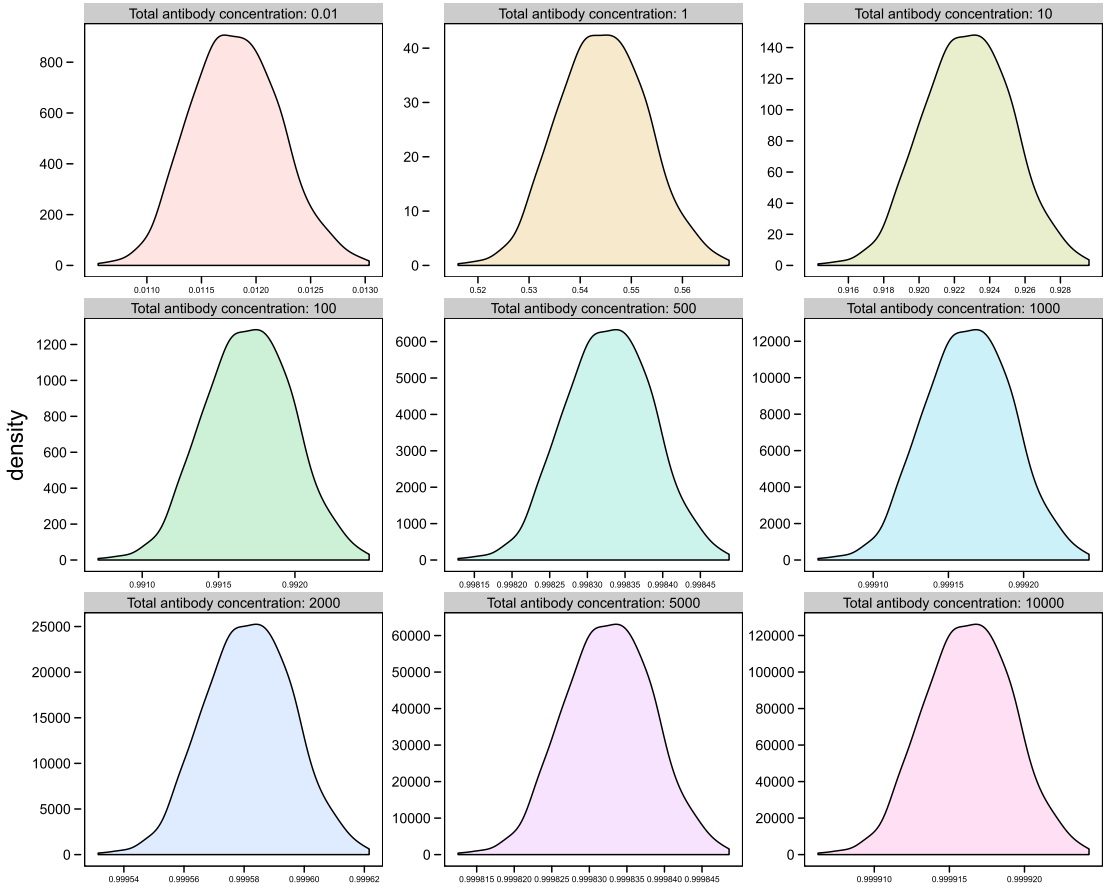


Figure 6.2: Gaussian kernel density estimates of signal intensities ($n_{\text{Ab}} = 10000$, unbiased mixture) in function of varying total antibody concentrations ($[\text{Ab}]_{\text{Total}} = 0.01$–$[\text{Ab}]_{\text{Total}} = 10000$) are shown. Colors are meant to improve readability. A random peptide library with associated AACM ($\mathbf{X}_{\text{sim}}$) of 1000 14-mers and assigned AAWS ($\vec{h}$) were generated once and were kept constant across all simulation runs. Antibody binding profiles were computed using Equation 4.1. Simulated signal intensities were not normalized.

6.3 Assessing the impact of total antibody concentration on the clustering of signal intensity profiles

The total antibody concentration does not only influence signal intensity values (Figures 6.2 and 6.3), but also the *Pearson* correlation structure of signal intensities due to its nonlinear contribution to signal intensity (Equations 4.1 and 6.1). Signal intensity profiles simulated with three different ranges of total antibody concentration were found to cluster by range both hierarchically (Figure 6.9) and variance-wise (Figure S.13). Additionally, signal intensity profiles could be classified by P-SVM nested cross-validation. Balanced accuracies range from 83.3% to 91.7% (Section A.7, Table S.3).

However, the clustering behavior differs with respect to the chosen method of correlation. Building the *Spearman* correlation matrix of the above simulated signal intensities

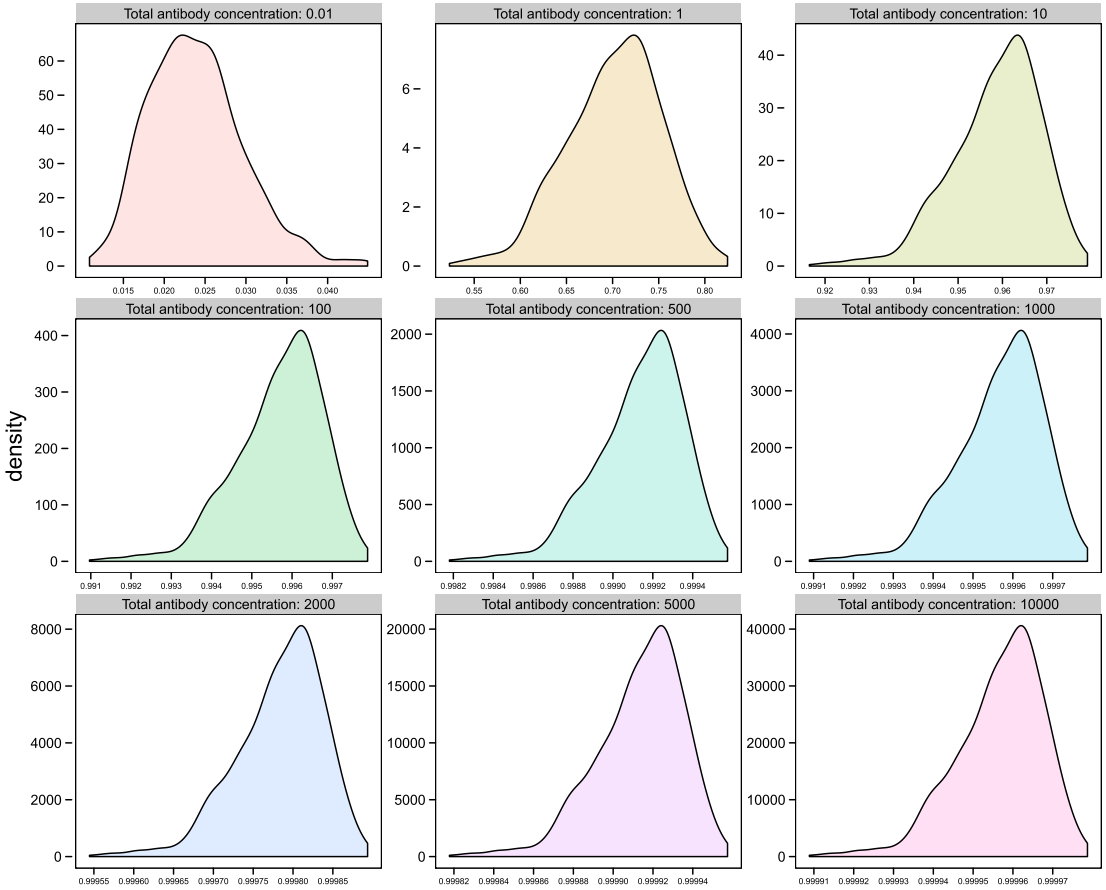


Figure 6.3: Gaussian kernel density estimates of signal intensity distributions ($n_{\text{Ab}} = 1$) in function of varying total antibody concentrations ($[\text{Ab}]_{\text{Total}} = 0.01$–$[\text{Ab}]_{\text{Total}} = 10000$) are shown. Colors are meant to improve readability. A random peptide library ($\mathbf{X}_{\text{sim}}$) with associated AACM of 1000 14-mers and assigned AAWS ($\vec{h}$) were generated once and were kept constant across all simulation runs. Simulations were done with one antibody ($n_{\text{Ab}} = 1$). Antibody binding profiles were computed using Equation 4.1. No normalization was applied to antibody binding profiles. The switching of density plots from positive skew for low $[\text{Ab}]_{\text{Total}}$ to negative skew for higher $[\text{Ab}]_{\text{Total}}$ is explained by Equation 4.1.

yields the identity matrix, which is due to the monotonic influence of total antibody concentration on signal intensity values (Figure 6.4) leaving ranks of signal intensities invariant (Section 6.2). The Spearman correlation is the Pearson correlation of the ranks of the input data (Section 3.9.1).

Thus, simulations suggest that *monotonic* changes in signal intensity profiles—induced by total antibody concentration differences—can be detected by Pearson correlation analysis. In addition, *non-monotonic* differences in signal intensity profiles such as rank changes—induced by antibody dominance (Section 4.3.4, Figures 4.3 and 7.1)—are detectable by Spearman correlation analysis.

Applying this knowledge to the Slovenian healthy study (SHS), the clustering of

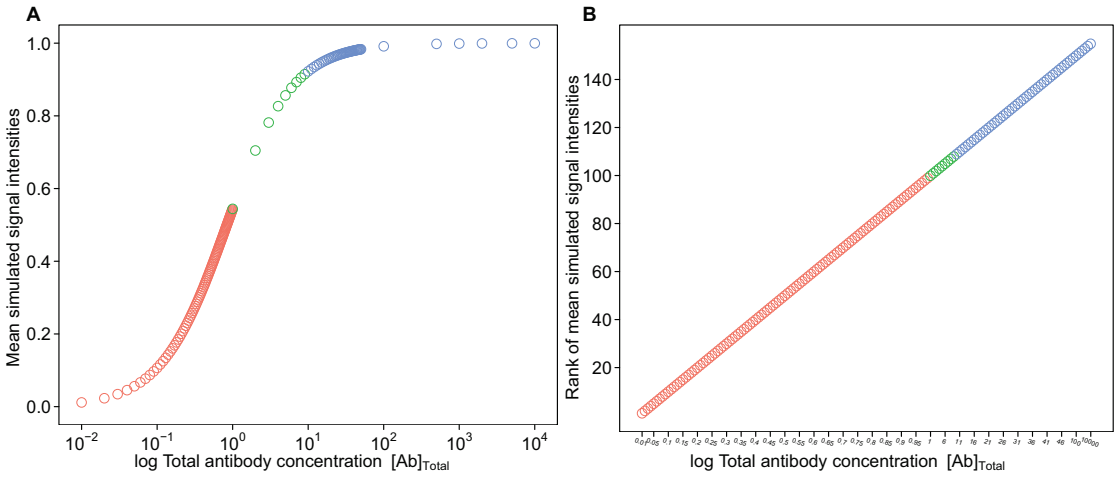


Figure 6.4: Mean simulated signal intensities vary in monotone increasing fashion with the total antibody concentration. (A) Mean signal intensities, obtained by averaging over all peptide signal intensities of a simulated peptide library ($\sum_{i=1}^{\#P} S_{\text{sim},i}$), as well as (B) their ranks are displayed in function of total antibody concentrations. Colors are meant to improve readability. Total antibody concentrations ($[\text{Ab}]_{\text{Total}}$) were varied between $[\text{Ab}]_{\text{Total}} = 0.01$ and $[\text{Ab}]_{\text{Total}} = 10000$. A random peptide library with associated AACM ($\mathbf{X}_{\text{sim}}$) of 1000 14-mers and assigned AAWS ($\vec{h}$) were generated once and were kept constant across all simulation runs. Simulations were performed with unbiased mixtures ($n_{\text{Ab}} = 10000$). Antibody binding profiles were computed using Equation 4.1. No normalization was applied to antibody binding profiles.

signal intensity profiles by healthy volunteer (Figures 6.8A–D) is suggested to be more likely due to dominant antibodies with volunteer-specific binding behavior than to total IgM concentration differences between healthy volunteers (Figure 6.7). In addition, the clustering of ranks by volunteer indicates that the putative dominant antibodies do not change over the sample collection period (1 month, Section 3.5.1).

6.4 Assessing the impact of the assigned AAWS distribution on signal intensity and predictive performance

In order to study in silico the influence of the distribution of assigned AAWS on simulated signal intensities and predictive performance, the components of assigned AAWS ($\vec{h}$) were set to 0 except for one component which was set to 1. Resulting signal intensity profiles for monoclonal antibodies (Figure 6.10A) and unbiased mixtures (Figure 6.10B) are multimodal. This multimodality is explained by examining the monoclonal case (Figure 6.10A); due to the chosen $\vec{h}$, most peptides ($\vec{p}^{\,i}$) will be equivalent to the null vector yielding a signal intensity (S_i) of 0.5. Therefore, for a total antibody concentration ($[\text{Ab}]_{\text{Total}}$) of 1, the kernel density estimates of the signal intensity profiles peak at 0.5 (Figure 6.10A). The remaining peaks are explained by peptides possessing at least one amino acid with a non-zero weight.

Of note, within the tested range of total antibody concentrations, predictive perfor-

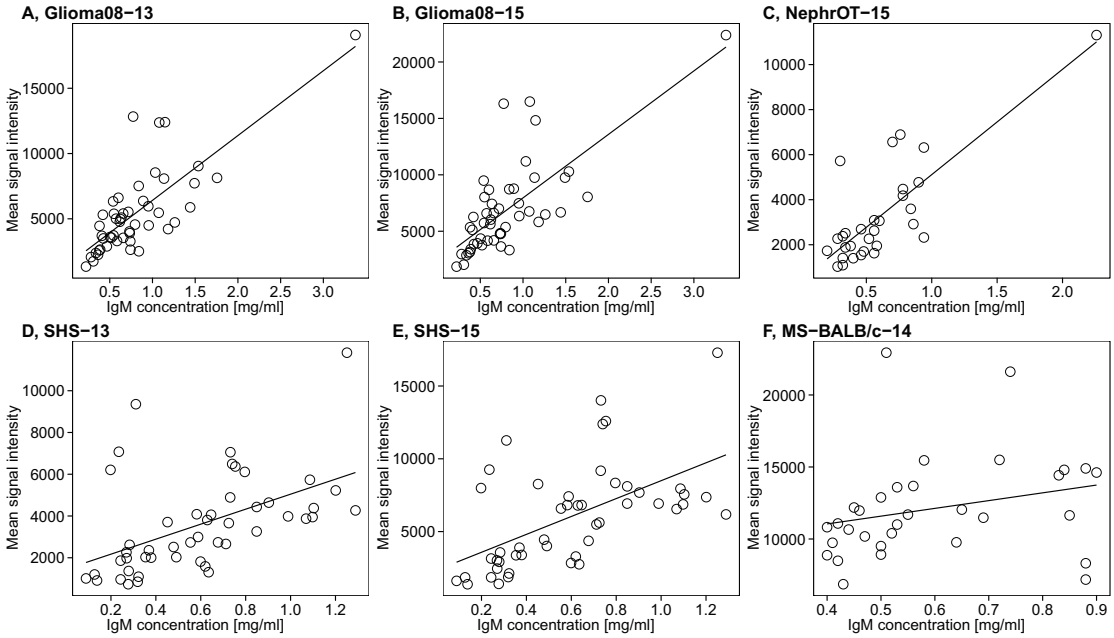


Figure 6.5: Assessment of the dependence of mean signal intensity on IgM concentration by experimental study. Corresponding correlation coefficients are displayed in Table 6.1. Signal intensities and IgM concentrations for the respective studies were measured as detailed in Section 3.5. The mean signal intensity was calculated by averaging over all non-normalized peptide signal intensities of an analyzed peptide library ($\sum_{i=1}^{\#P} S_i$). Subplots are named according to the convention: Experimental study-(Array manufacturer/Mouse model)-Peptide length. Experimental studies displayed are those for which the serum or plasma was diluted 1:10 and IgM concentration data were available (Section 3.5).

mance is not affected by a change in assigned AAWS $\vec{h}$ for neither unbiased mixtures nor monoclonal antibodies (data not shown as Figures are analogous to Figure 6.6)

6.5 Violating the assumption of the random generation of antibody sequences decreases predictive performance

Previously, antibody sequences were simulated in random, i.i.d.-fashion, such that they were no more than randomly correlated (Section 4.2). In order to simulate correlated antibody mixtures, increasing Gaussian noise (until $\sigma = 10$) was added to 40 different antibody sequences (Section 3.6.3, Figure 3.2). Thereby, antibody mixtures of 10000 antibodies each were built. Figure 6.11 shows that the predictive performance as well as the recovery of assigned AAWS $\vec{h}$ is low when the correlation between antibody sequences is high (Table 3.4). This is similar to low-diversity antibody mixtures (Figures 4.2A and C). With increasing noise amplitude—thereby increasing the decorrelation[1] of simulated antibody mixtures (Table 3.4)—predictive performance values and recovery of assigned

[1]The explanation of the term "decorrelation" is found in Section 3.6.3.

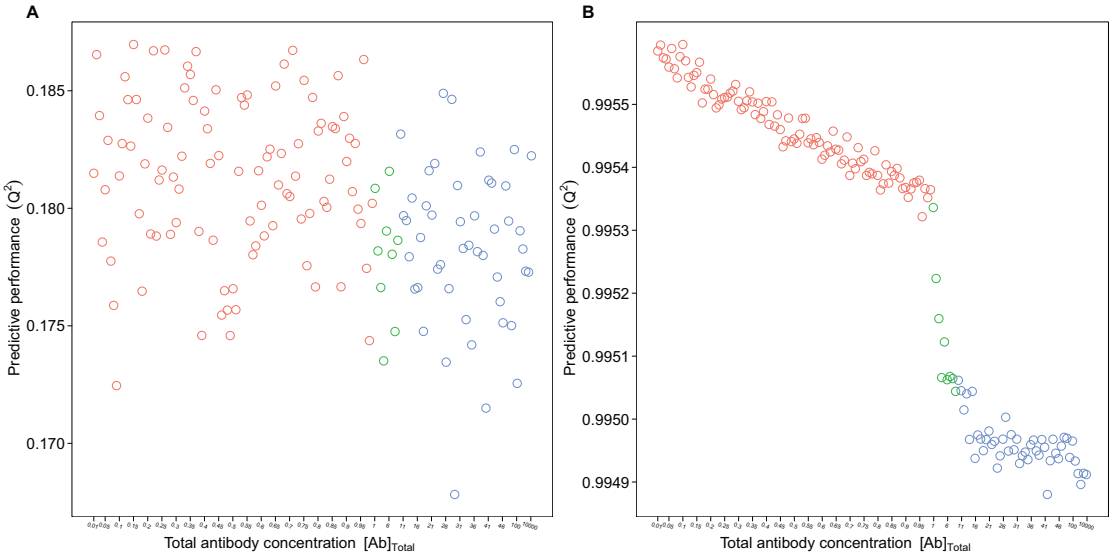


Figure 6.6: Simulations show that the predictive performance is marginally impacted by a varying total antibody concentration ($[Ab]_{Total}$ =0.01–10000) for both (A) monoclonal antibodies ($n_{Ab} = 1$, $r_{Pearson} = -0.20$, $r_{Spearman} = -0.36$) and (B) unbiased mixtures ($n_{Ab} = 10000$, $r_{Pearson} = -0.21$, $r_{Spearman} = -0.98$). Colors are meant to improve readability. A random peptide library with associated AACM ($\mathbf{X}_{sim}$) of 1000 14-mers and assigned AAWS ($\vec{h}$) were generated once and were kept constant across all simulation runs.

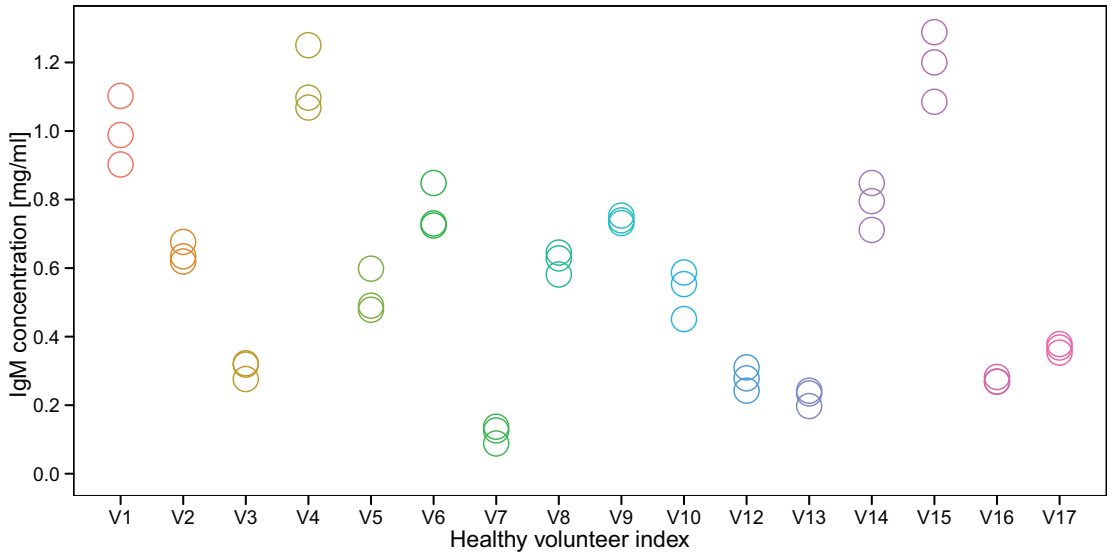


Figure 6.7: Slovenian healthy study: IgM concentrations show low intra but high inter-volunteer differences with up to a tenfold differences between volunteers (e. g. V7 compared to V4). Number of samples: 48, 3 samples for each of the 16 healthy volunteers (V*, Section 3.5.1).

AAWS $\vec{h}$ improve (Figure 6.11), eventually reaching values similar to those of unbiased mixtures (Figures 4.2A and C).

Experimental study	Correlation: [IgM] vs. Mean SI			
	r_{Pearson}	p_{Pearson}	r_{Spearman}	p_{Spearman}
Glioma08-13	0.77	0.00	0.71	0.00
Glioma08-15	0.73	0.00	0.71	0.00
NephrOT-15	0.80	0.00	0.65	0.00
SHS-13	0.50	0.00	0.60	0.00
SHS-15	0.56	0.00	0.60	0.00
MS-BALB/c-14	0.26	0.15	0.41	0.02
	Correlation: [IgM] vs. Q^2			
Glioma08-13	-0.11	0.45	-0.16	0.28
Glioma08-15	0.08	0.56	-0.03	0.85
NephrOT-15	0.24	0.20	0.04	0.84
SHS-13	-0.12	0.43	-0.10	0.49
SHS-15	-0.07	0.63	-0.14	0.33
MS-BALB/c-14	-0.13	0.48	-0.08	0.67

Table 6.1: Correlation coefficients and corresponding p-values between IgM concentration ([IgM]) and both mean signal intensity (Mean SI) and predictive performance values (Q^2) are tabled by experimental study. The mean signal intensity ($\sum_{i=1}^{\#P} S_i$) was calculated by averaging over a peptide library's non-normalized signal intensity profile. Signal intensities and IgM concentrations for the respective studies were measured as detailed in Section 3.5. Experimental studies are named according to the convention: Experimental study-(Mouse model)-Peptide length.

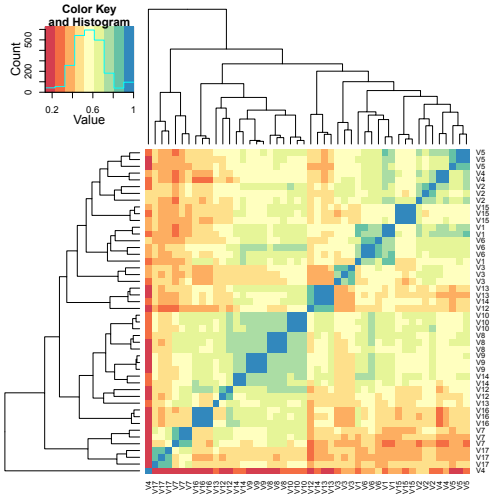
Thus, the violation of the i.i.d.-assumption of antibody sequence generation negatively impacts predictive performance. This impact differs with respect to monoclonal antibody strength[2]. Therefore, not only an antibody mixture's diversity but also its composition have an impact on predictive performance: mixtures which are both highly diverse and highly correlated (Figure 6.11A) yield predictive performance values similar to those yielded by low-diversity random mixtures (Figure 4.2A).

6.6 Summary

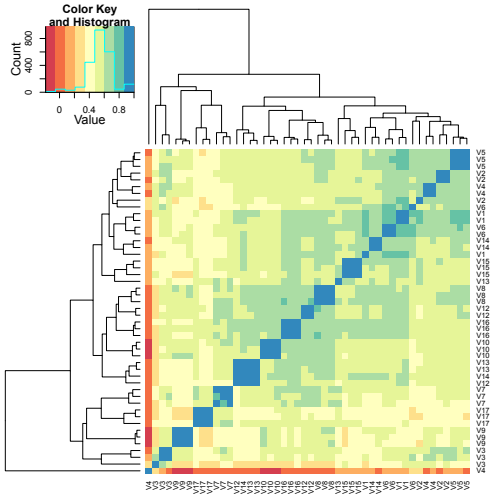
- The predictive performance and recovery of assigned AAWS of unbiased mixtures are robust against variations in peptide library size and peptide length. For the tested ranges of both parameters, the median predictive performance remains well above $Q^2 = 0.95$. However, increased peptide length slightly decreases predictive performance and higher peptide library sizes result in a slight increase of both predictive performance and recovery of assigned AAWS (Section 6.1, Figure 6.1).

- A growing total antibody concentration leads to the saturation of simulated signal intensities ($\vec{S} \to \vec{1}$) whereas if total antibody concentrations tend to 0, simulated signal intensities will tend to 0 ($\vec{S} \to \vec{0}$, Section 6.2, Figures 6.2 and 6.3). In

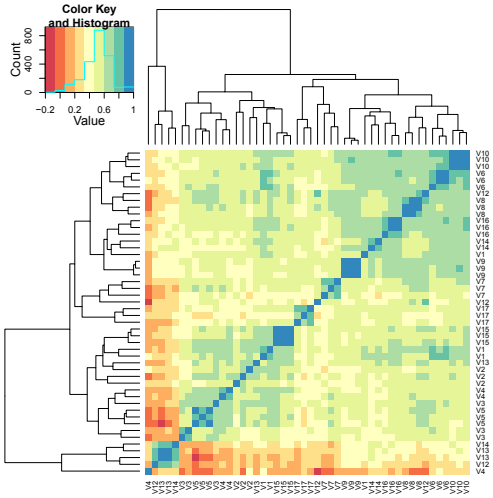
[2]For further information on antibody strength, please refer to Section 7.2.



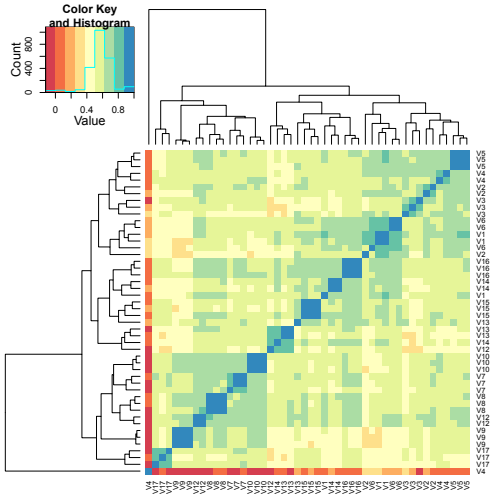
6.8A: Heatmap of SI, 13-mers



6.8B: Heatmap of SI, 15-mers



6.8C: Heatmap of ranks, 13-mers



6.8D: Heatmap of ranks, 15-mers

Figure 6.8: Slovenian healthy study: antibody binding profiles and their ranks cluster mostly by healthy volunteer. The heatmaps (Section 3.9.2) of 13- and 15-mer IgM signal intensity profiles (A and C) and their ranks (B and D) are shown. Heatmaps were built based on the Pearson correlation matrix of non-normalized signal intensities and ranks. Hierarchical clustering recovers (A) 11, (B) 10, (C) 11, (D) 10 out of 16 triplets, where "triplet" denotes the three samples obtained from each healthy volunteer (Section 3.5.1). IgM signal intensity profiles were measured with the peptide libraries of $J^{2304}_{\text{13-mer}}$ and $J^{3418}_{\text{15-mer}}$ peptides for 13- and 15-mers, respectively. Number of serum samples: 48, 3 samples for each of the 16 healthy volunteers (V*).

agreement with simulation results (Figure 6.4), in vitro data show for most tested experimental studies a significant dependence of mean signal intensity on total IgM

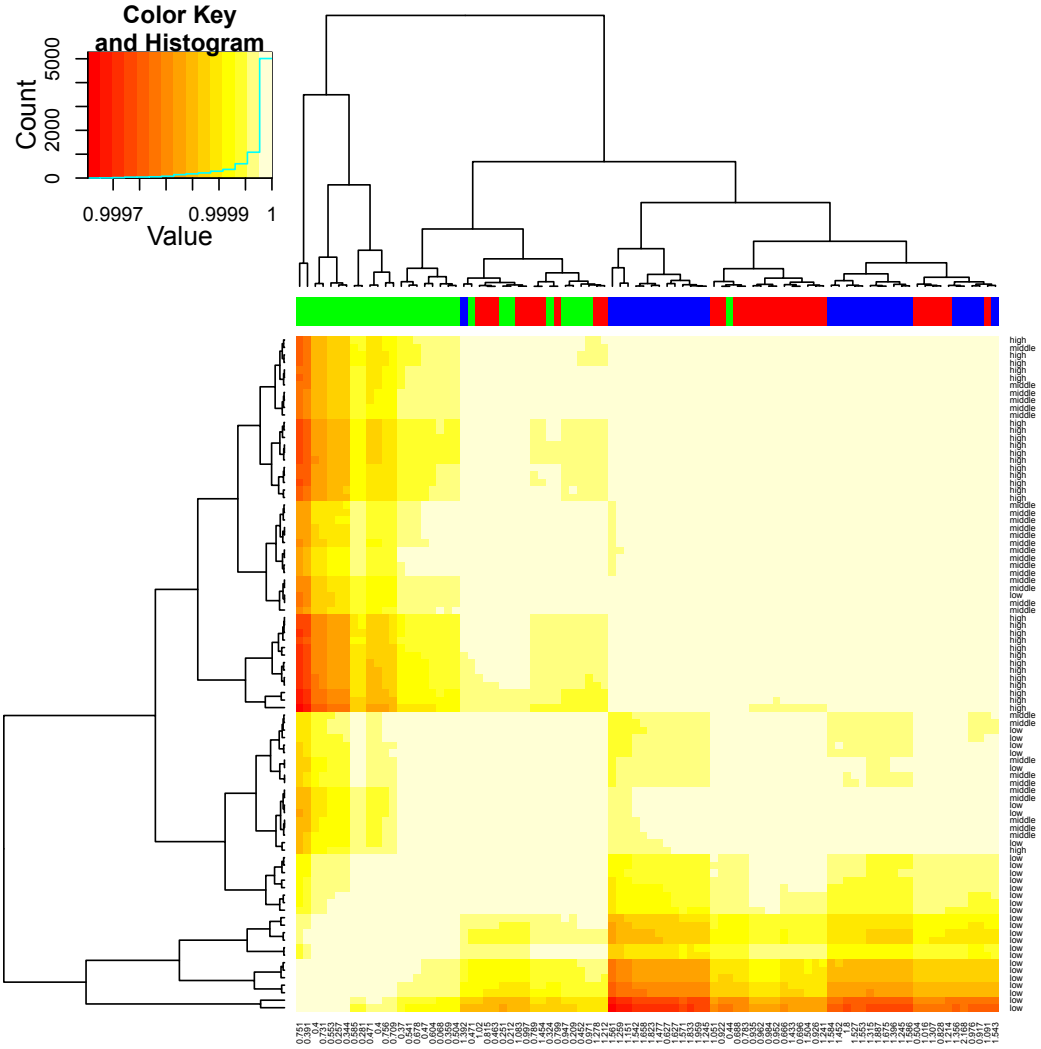


Figure 6.9: Simulations show that total antibody concentration differences have an impact on Pearson-based hierarchical clustering of antibody binding profiles of unbiased mixtures ($n_{\mathrm{Ab}} = 10000$). Thirty different total antibody concentrations ($[\mathrm{Ab}]_{\mathrm{Total}}$) were randomly drawn from each of the three different Gaussian distributions $\mathcal{N}(0.5, 0.25)$, $\mathcal{N}(1, 0.25)$ and $\mathcal{N}(1.5, 0.25)$ [shown in the column labels of the heatmap] termed "low" (green), "middle" (red), and "high" (blue) [row labels], respectively, to simulate antibody binding profiles (Equation 4.1). Subsequently, non-normalized signal intensity profiles (255 random 14-mers) were subjected to Pearson-based hierarchical clustering (Section 3.9.2). Across all simulation runs, the unbiased mixture was held constant. PCA of the simulated signal intensity profiles shows clustering as well (Figure S.13).

concentration (Figure 6.5, Table 6.1).

- Predictive performance values are marginally influenced by the tested range of total antibody concentrations both in simulations (Section 6.2, Figure 6.6) and in vitro (Table 6.1).

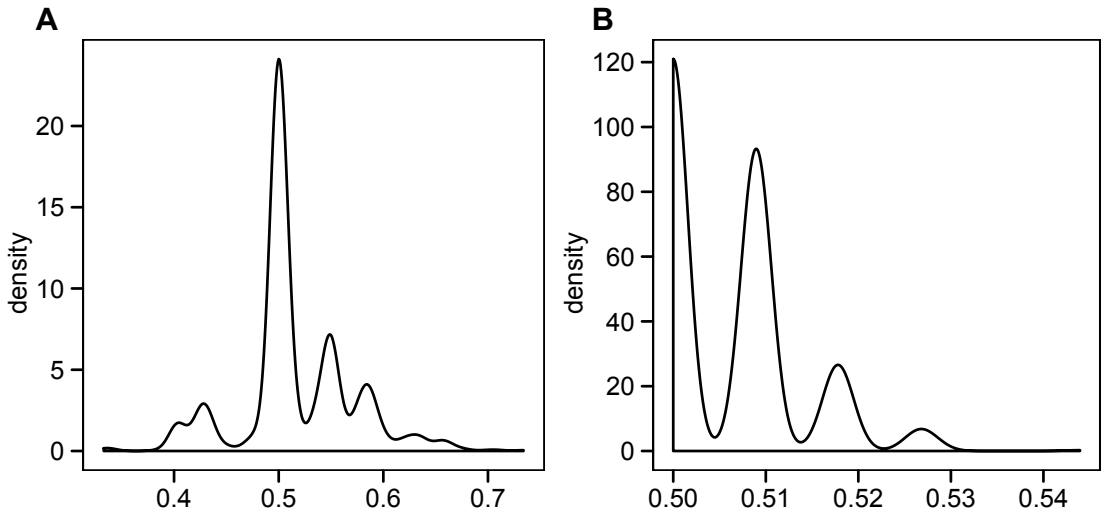


Figure 6.10: Gaussian kernel density estimates of signal intensity profiles are dependent on assigned AAWS. A simulated random peptide library with associated AACM ($\mathbf{X}_{\text{sim}}$) of 1000 14-mers and assigned AAWS ($\vec{h} = (0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1)^{\text{T}}$) were generated once and were kept constant across all simulation runs. Simulations were done with (A) one antibody ($n_{\text{Ab}} = 1$, $[\text{Ab}]_{\text{Total}} = 1$) and (B) an unbiased mixture ($n_{\text{Ab}} = 10000$, $[\text{Ab}]_{\text{Total}} = 1$). Antibody binding profiles were computed using Equation 4.1.

- Due to its nonlinear and monotonic increasing impact on simulated signal intensities (Section 6.2, Figure 6.4), the total antibody concentration has an influence on the clustering of simulated signal intensity profiles in case of Pearson correlated signal intensity profiles (Figures 6.9 and S.13, Table S.3). No influence on Spearman correlation of simulated profiles was evident (Section 6.3). Consequently, experimentally determined signal intensity profiles of samples from the Slovenian healthy study are suggested to cluster rather due to relative antibody mixture differences (e.g. different rank-altering antibody dominances, Figure 4.3) than due to differences in total IgM concentrations (Figures 6.7 and 6.8).

- The distribution chosen for assigned AAWS $\vec{h}$ influences the modality of signal intensity distributions (Section 6.4, Figure 6.10). However, the studied distribution change has no impact on predictive performance.

- Simulations show that a violation of the i.i.d.-assumption of antibody sequence generation decreases predictive performance and recovery of assigned AAWS (Section 6.5, Figure 6.11). Both an antibody mixture's diversity and composition have an impact on predictive performance: mixtures which are both highly diverse and highly correlated (Figure 6.11A) yield predictive performance values similar to those yielded by low-diversity random mixtures (Figure 4.2A).

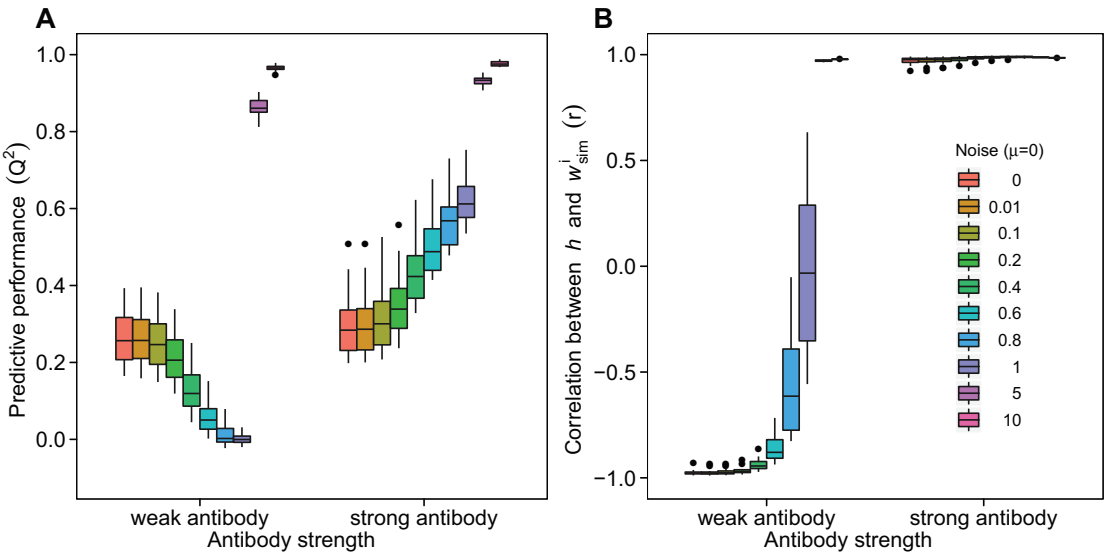


Figure 6.11: Simulations show that highly correlated antibody mixtures yield low predictive performance values. (A) The predictive performance (Q^2) of antibody mixtures which are correlated to varying degrees is shown. (B) The correlation of estimated AAWS ($\vec{w}^i_{\text{sim}}$) with the assigned AAWS $\vec{h}$ is shown. A random peptide library with associated AACM ($\mathbf{X}_{\text{sim}}$) of 1000 14-mers and assigned AAWS ($\vec{h}$) were generated once and were kept constant across all simulation runs. Antibody binding profiles were computed using Equation 4.1. Corresponding AAWS ($\vec{w}^i_{\text{sim}}$) were determined using Equation 4.8. The 20 "strongest" (highest $\langle y_{i,k} \rangle = \langle (a^k)^{\mathrm{T}} p^i \rangle$) and "weakest" monoclonal antibodies (lowest $\langle y_{i,k} \rangle = \langle (a^k)^{\mathrm{T}} p^i \rangle$) of the 500 monoclonal antibodies used in Figure 7.1A were chosen for this simulation. Based on the chosen 40 antibodies, antibody mixtures of 10000 antibodies with varying degree of correlation were generated (Section 3.6.3, Figure 3.2, Table 3.4). The Gaussian noise term used had the following parameters: $\mathcal{N}(\mu = 0, \sigma = x)$ with x ranging from 0 to 10.

7 A minimal model of antibody-peptide binding: monoclonal antibodies

7.1 Signal intensity profiles as well as AAWS of simulated monoclonal antibodies are isotropically distributed in the variance space

Simulations show that signal intensity profiles as well as AAWS of 500 randomly generated antibodies are isotropically distributed in the space spanned by the first two principal components (Figures 7.1A and 7.1B). Thus, variance-wise, clustering of both simulated signal intensities and AAWS of biased antibody mixtures is only possible if the biasing dominant antibodies have similar binding properties with respect to the peptide library studied (Figures 7.1C and 7.1D).

7.2 Simulated monoclonal antibodies can be separated into two groups based on their performance to recover assigned AAWS

Simulated monoclonal antibodies can be separated into two different groups according to a criterion termed "antibody strength". Strong antibodies are defined as yielding a mean binding association ($\langle y_{i,k} \rangle_{|k=\text{const}} = \langle (\vec{a}^k)^{\mathrm{T}} \vec{p}^{\,i} \rangle = \sum_i \sum_z a_z^k p_z^i$) higher than 0.5. Otherwise, antibodies are termed "weak".

AAWS, but not signal intensity profiles, of weak and strong antibodies can be separated by PCA (Figures 7.2A and B)[1]. Strong antibodies cluster in the variance space spanned by PCA because they are—in contrast to weak antibodies—highly correlated to assigned AAWS (Figure 7.2I). Sequences of strong antibodies show a higher number of positive components, higher mean signal intensities[2] and slightly higher predictive performance values (Figures 7.2D, E and H).

Strong antibodies recover assigned AAWS $\vec{h}$ well because they have a high number of positive components. In fact, both the number of positive components (Figure 7.2D) and the components' means (Figure 7.2E) are highly correlated to the recovery of assigned AAWS $\vec{h}$ (Figure 7.2I, $r_{\text{Pearson, Spearman}} > 0.84$). The more the number of positive and

[1] Higher order principal components do not show any separation of signal intensities by antibody strength either (data not shown).

[2] This observation is only possible if signal intensities were neither log-transformed nor set to zero mean and unit variance.

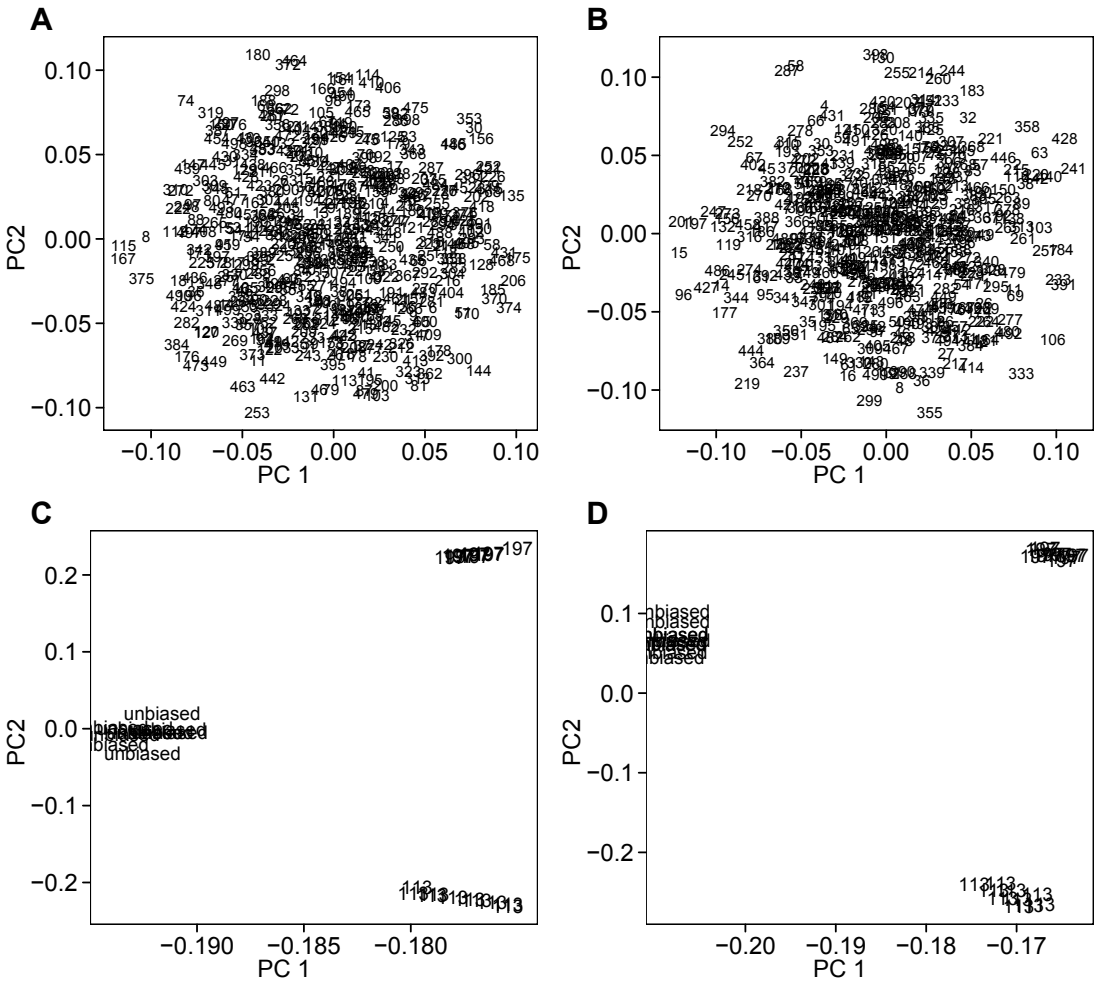


Figure 7.1: Normalized simulated signal intensity profiles and AAWS of biased mixtures separate in the space spanned by the first two principal components if dominant antibodies show similar binding properties with respect to the peptide library used. (A) Normalized signal intensity profiles and (B) AAWS of 500 random simulated monoclonal antibodies are shown in the two-dimensional space spanned by the first two principal components (PC1, PC2). (C) Ten signal intensity profiles of biased mixtures of 10000 antibodies in which one dominant antibody, labeled "113" or "197", is increased 1000-fold in concentration were simulated. The dominant antibodies were kept constant across simulation runs, whereas the other 9999 antibodies were varied for each of the ten runs. In addition, ten signal intensity profiles of an unbiased mixture of 10000 antibodies (labeled as "unbiased") were computed. The 30 (10+10+10) signal intensity profiles are shown in the space spanned by first two principal components as are (D) their corresponding AAWS. The first two principal components explain 21.2% (A), 62.7% (B) , 95.7% (C) and 99.7% (D) of the variance in the data, respectively. In (A–D) a random peptide library with associated AACM ($\mathbf{X}_{\text{sim}}$) of 255 14-mers and assigned AAWS ($\vec{h}$) were generated once and were kept constant across all simulation runs. Antibody binding profiles were computed using Equation 4.1. Corresponding AAWS ($\vec{w}^{i}_{\text{sim}}$) were determined using Equation 4.8.

negative components is equilibrated in an antibody sequence (e.g. for $l = 14$: 7 negative and 7 positive components) or the more the components' means tend to zero, the more likely it is that the recovery of assigned AAWS tends to zero ($r \to 0$). In contrast, the more antibody sequences are biased toward positive or negative components, or the higher or lower the components' means are, the more likely it is that the recovery of assigned AAWS tends to 1 or -1, respectively.

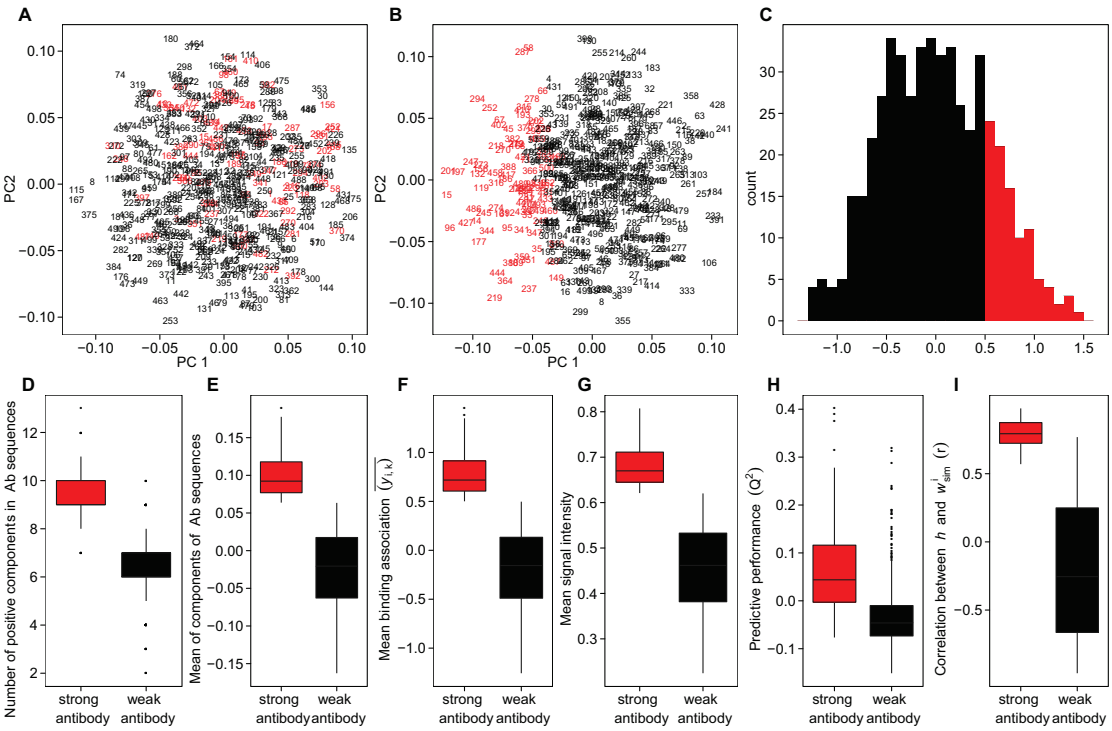


Figure 7.2: Strong antibodies show a better recovery of assigned AAWS $\vec{h}$ compared to weak antibodies. (A) Normalized signal intensity profiles and (B) AAWS of the 500 simulated monoclonal antibodies of Figure 7.1 are shown in the space spanned by the first two principal components (PC1, PC2). Strong antibodies, shown in red, are defined as having a mean binding association ($\langle y_{i,k} \rangle$) higher than 0.5. (C) A histogram of the 500 mean binding associations ($\langle y_{i,k} \rangle$) shows that there are more weak ($n = 403$) than strong antibodies ($n = 97$). (D) Compared to weak antibodies, strong antibodies have a higher number of positive components, (E) which results in higher component means, (F) per definitionem higher mean binding associations, (G) higher mean non-normalized signal intensities ($\sum_{i=1}^{\#P} S_{sim,i}$), (H) slightly higher predictive performance values (Q^2), as well as (I) a better recovery of assigned AAWS $\vec{h}$. Differences between boxplots are significant ($p < 0.05$).

7.3 The criterion of antibody strength is robust against peptide library changes but not against changes in assigned AAWS

The binding association $y_{i,k}$ (Equation 4.1) is determined by two variables: (i) antibody and (ii) peptide sequence. It is therefore of interest to determine, whether the criterion of antibody strength depends on the given peptide library.

Simulations show that—assigned AAWS being constant—a change of peptide library alters marginally the subset of strong antibodies (Figure 7.3A). This is due to the random generation of peptide libraries which results in a similar *mean* binding association per antibody[3] ($\langle y_{i,k} \rangle$).

In contrast, antibody strength is sensitive to varying assigned AAWS $\vec{h}$ (Figures 7.3B and C): the number of strong antibodies found for a given peptide library is highly correlated to the mean of the components of given assigned AAWS ($r = 0.99$). Thus, assigned AAWS modulate the number of strong and weak antibodies found for a given peptide library.

7.4 Assessment of the in vitro evidence for antibody strength

The assessment of antibody strength in vitro can only be done indirectly. This is so because assigned AAWS of the used peptide libraries are not known a priori which renders the assessment of the quality of their recovery, one of the primary defining features of strong antibodies (Figure 7.2I), unattainable. Simulations (Figure 4.2) predict that the higher the predictive performance of an antibody mixture (or a monoclonal antibody), the better the recovery of assigned AAWS $\vec{h}$ should be.

Assuming that estimated AAWS ($\vec{w}$) of healthy BALB/c mice (Section 5.1) near assigned AAWS $\vec{h}$ of the $J^{255}_{14\text{-mer}}$-library, simulations predict that the Pearson correlation between estimated AAWS ($\vec{w}$) of monoclonal antibodies and estimated AAWS of healthy BALB/c mice *increases* with increasing predictive performance of monoclonal antibodies[4]. Indeed, predictive performance of monoclonal antibodies *positively* correlates ($r_{\text{Pearson}} = 0.86$, $r_{\text{Spearman}} = 0.89$)[5] with the correlation between AAWS of monoclonal antibodies and AAWS of healthy BALB/c mice (Figures 7.4A and 7.4C).

However, mean raw signal intensities of monoclonal antibodies[6] only Pearson/Spearman correlate poorly with predictive performance ($r_{\text{Pearson}} = 0.22$, $r_{\text{Spearman}} = 0.32$) or the median correlation between estimated AAWS of monoclonal and BALB/c serum antibodies ($r_{\text{Pearson}} = 0.23$, $r_{\text{Spearman}} = 0.45$), which is in contrast to simulation results

[3] Antibody strength is only independent of the given peptide library, assuming assigned AAWS are kept constant, up until a certain minimal library size (data not shown).

[4] Both monoclonal antibodies and sera of BALB/c mice were incubated on the $J^{255}_{14\text{-mer}}$ library (Table 3.1).

[5] Both correlation coefficients are significant ($p < 0.01$).

[6] Normalization of monoclonal antibody binding profiles was not performed for the here described correlation as it would have rendered the analysis pointless. The high impact of detection antibodies on raw signal intensity profiles (Figure S.4), however, could have biased the performed correlation analysis.

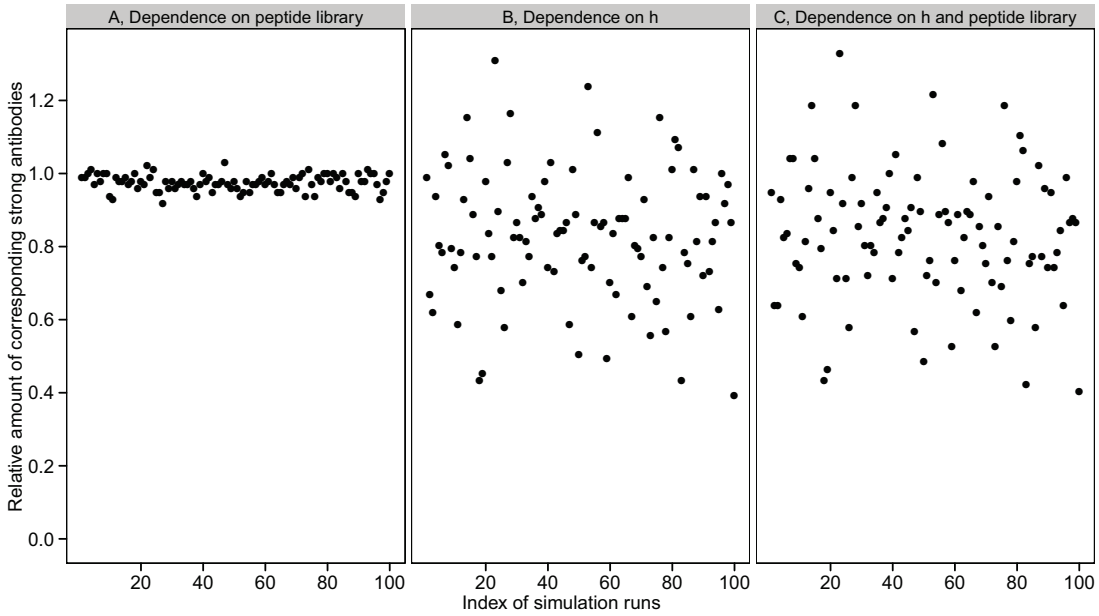


Figure 7.3: The criterion of antibody strength, which divides simulated antibodies into weak and strong antibodies, is robust against changing peptide libraries but sensitive to varying assigned AAWS $\vec{h}$. The strength of the 500 monoclonal antibodies of Figure 7.1 was determined with respect to a library of 255 14-mers of which either (A) the composition, (B) the assigned AAWS ($\vec{h}$) or (C) both the composition and assigned AAWS ($\vec{h}$) were altered. For each case (A, B, C) 100 simulations were run. In (A–C) for each simulation run, the strong antibodies were determined, compared for correspondence with the strong antibodies determined for the peptide library and assigned AAWS used in Figure 7.1. Subsequently, the number of corresponding strong antibodies was divided by the total number of strong antibodies of Figure 7.1 thus forming the ratio which constitutes the y-axis. This ratio can, by definition, exceed 1.

(Figures 7.2G–I). Summarizing, in vitro results are in accord with simulation results in that monoclonal antibodies which have higher predictive performance values are also higher correlated to AAWS of highly diverse antibody mixtures. However, the in silico concept of antibody strength is only partly supported by in vitro antibody-peptide reactivity data.

Of note, the predictive performance of monoclonal antibodies correlates *negatively* with the correlation between AAWS from monoclonal antibodies and AAWS from healthy individuals from the SHS ($r_{\text{Pearson}} = -0.48$, $r_{\text{Spearman}} = -0.54$)[7]. In addition, whereas estimated AAWS of monoclonal antibodies "VB142" and "eiJB40" correlate highest with AAWS of BALB/c sera (Figure 7.4A), estimated AAWS of the monoclonal antibody "VB176" correlate highest with AAWS of human sera (Figure 7.4B).

[7]Both correlation coefficients are not significant $0.05 < p_{\text{Pearson}}, p_{\text{Spearman}} < 0.10$.

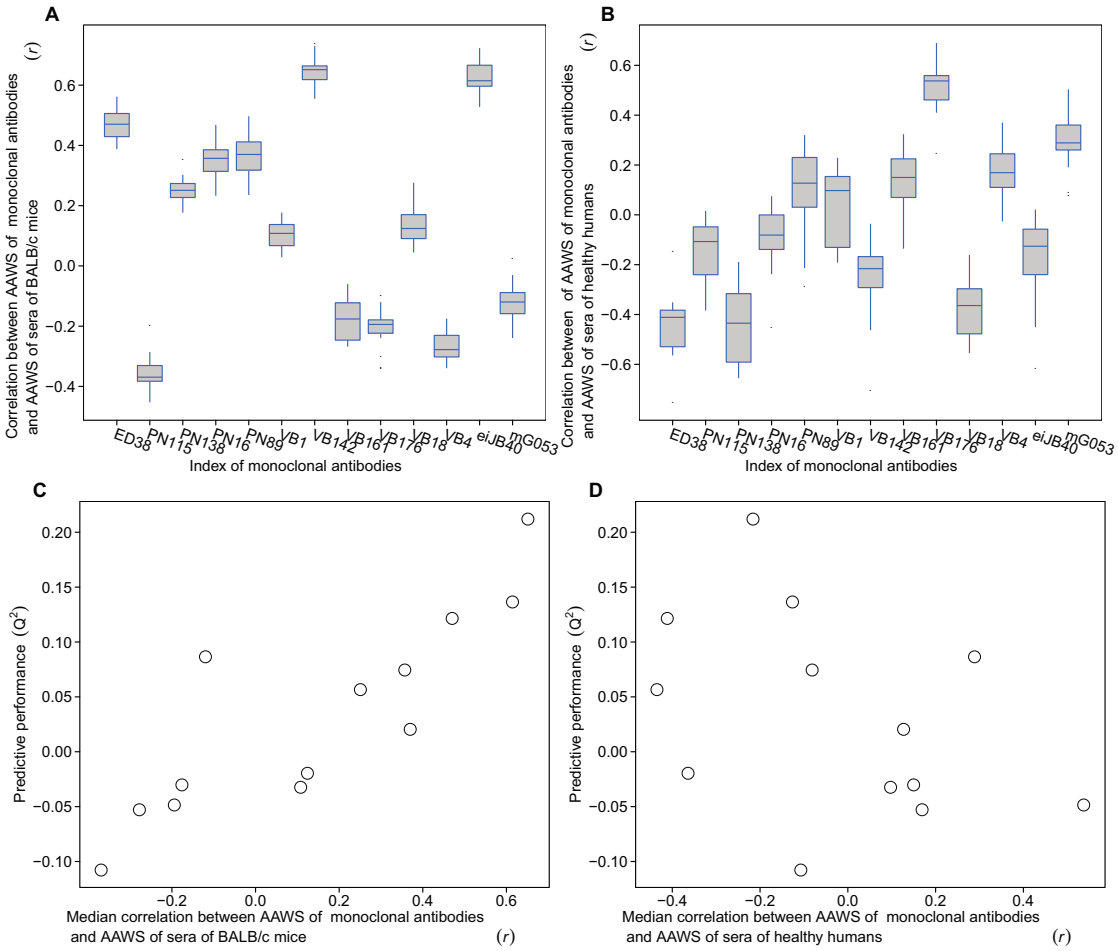


Figure 7.4: Predictive performance of monoclonal antibodies correlates positively with the median Pearson correlation between AAWS of monoclonal antibodies and AAWS of healthy BALB/c mice. The Pearson correlation of estimated AAWS of 13 monoclonal antibodies with both (A) 15 AAWS of sera of healthy BALB/c mice (MS) and (B) 48 AAWS of human samples (SHS) are shown. (C, D) The median correlation coefficients of each boxplot in (A, B) were determined and plotted in function of the predictive performance values of monoclonal antibodies (C: $r_{\text{Pearson}} = 0.86$, $r_{\text{Spearman}} = 0.89$, $p_{\text{Pearson, Spearman}} < 0.01$, D: $r_{\text{Pearson}} = -0.48$, $r_{\text{Spearman}} = -0.54$, $0.05 < p_{\text{Pearson, Spearman}} < 0.10$). Antibody binding profiles of monoclonal antibodies and sera were determined as detailed in Sections 3.5.1, 3.5.8 and 3.5.9.

7.5 Antibody strength impacts antibody binding profiles of correlated antibody repertoires

Simulations show that the predictive performance of correlated antibody repertoires is dependent on antibody strength (Figure 6.11). For antibody repertoires based on strong antibodies, an increasing decorrelation (Table 3.4) leads to higher predictive performance values of generated antibody repertoires nearing perfection for a noise amplitude of

$\mathcal{N}(\mu = 0, \sigma = 10)$. This is partly in contrast to the behavior of correlated antibody repertoires of weak antibodies for which up to a certain noise amplitude, predictive performance does not increase. However, for high noise amplitudes, ($\mathcal{N}(\mu = 0, \sigma = 5)$ and $\mathcal{N}(\mu = 0, \sigma = 10)$), predictive performance values of antibody repertoires generated on the basis of weak antibodies are analogous to those based on strong antibodies. The explanation for the difference in predictive performance values between weak and strong antibodies for low and medium noise terms is similar to that for the weak antibodies' poor recovery of assigned AAWS (Section 7.2). Weak antibody sequences are biased to possessing negative components (Figure 7.2D). Adding Gaussian noise terms to these antibody sequences results in a transient equilibration of positive and negative components which was suggested (Section 7.2) to be the cause of a poor recovery of AAWS.

7.6 Summary

- Signal intensity profiles and AAWS of random simulated monoclonal antibodies are isotropically distributed in the variance space. Therefore, clustering of simulated signal intensity profiles and AAWS in the space spanned by PCA is dependent on dominant antibodies which have similar binding patterns with respect to a given peptide library (Section 7.1, Figure 7.1).
- Within the variance space spanned by PCA, AAWS of a subset of simulated monoclonal antibodies—strong antibodies—cluster but not their corresponding signal intensity profiles (Figure 7.2A and B).
- Strong antibodies—defined as showing a mean binding association ($\langle y_{i,k} \rangle$) higher than 0.5— differ from weak antibodies by having a higher number of positive components, higher non-normalized mean signal intensities, slightly higher predictive performance values and an enhanced recovery of assigned AAWS $\vec{h}$ (Section 7.2, Figure 7.2).
- The criterion of antibody strength is robust against peptide library changes but sensitive to varying assigned AAWS (Section 7.3, Figure 7.3).
- The in silico concept of antibody strength is only partly supported by in vitro antibody-peptide reactivity data (Section 7.4, Figures 7.4 and 7.2G–I).
- The predictive performance of monoclonal antibodies correlates positively with the correlation between AAWS of monoclonal antibodies and AAWS of healthy BALB/c mice (Section 7.4, Figures 7.4A and 7.4C). This in accord with predictions made by the mathematical model (Figure 4.2).
- The correlation structure of AAWS of monoclonal antibodies with AAWS of healthy BALB/c mice (Figures 7.4A and 7.4C) differs from that observed with AAWS of human individuals (Figures 7.4B and 7.4D).
- Correlated antibody repertoires behave differently regarding predictive performance depending on the antibody strength of the monoclonal antibody they were generated from (Section 7.5, Figure 6.11).

8 Technological analysis of antibody-peptide reactivity data

Parts of this Chapter were recently published [104].

The predictive performance of AAWS estimated from in vitro measurements of sera is not perfect (Section 5.2, Figure 5.1). This could at least be in part due to noise in signal intensity measurements. In the following, it is attempted to study with the help of simulations the influence of noise on predictive performance and recovery of assigned AAWS in function of peptide library parameters (Sections 8.1 and 8.2). Only simulations of *unbiased* mixtures are shown in this Chapter.

The formulation of the mathematical model of antibody-binding (Section 4.2) makes no assumptions about the dependence of estimated AAWS on genetic background, species, microarray batch and manufacturer. In this Chapter, the impact of these in vitro parameters on AAWS estimation is assessed (Section 8.3).

8.1 Simulations show that the introduction of noise into signal intensities decreases both the predictive performance and the recovery of assigned AAWS

Introducing Gaussian noise into simulated signal intensities (Section 3.6.2, Figure 8.4) of unbiased mixtures results, depending on the noise amplitude, in a decrease of predictive performance and recovery of assigned AAWS $\vec{h}$ by estimated AAWS $\vec{w}$ (Figure 8.1). Of note, even for noise amplitudes ($\mathcal{N}(\mu = 0, \sigma = 0.03)$, Figure 8.1A) for which predictive performance nears zero, the median Pearson correlation of assigned AAWS and estimated AAWS is about $r = 0.7$ (Figure 8.1C).

Thus, even though Gaussian noise destroys the linear relationship between the assigned AAWS (the AACM) and the signal intensity profile, both the recovery of assigned AAWS by estimated AAWS (Figure 8.1C) and the recovery of original signal intensities[1] ($\hat{S}_{\text{recovered}} = \mathbf{X}\vec{w}$, Figure 8.2) do not fall much below a median correlation coefficient of 0.5. Indeed, across all noise amplitudes, recovered signal intensities are higher correlated to original signal intensities than to noise-altered ones (Figure 8.2).

This property of PLSR being able to separate, up to a certain point, noise from signal is owed to its ability to migrate the noise terms into the higher components of the PLSR model. This PLSR property is exemplified in Figure 8.3; the higher the noise amplitude is increased, the less components are significantly predictive and therefore used by the model.

[1] Original signal intensities: signal intensities prior to introduction of Gaussian noise.

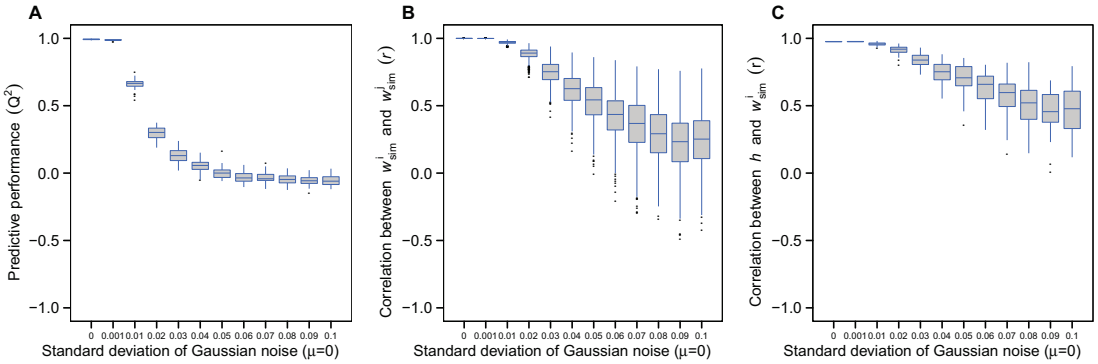


Figure 8.1: Simulations show that the predictive performance of antibody binding profiles of unbiased mixtures ($n_{\text{Ab}} = 10000$) decreases with increasing introduction of Gaussian noise. (A) Predictive performance (Q^2) decreases with increasing noise, (B) as does the correlation (r) between all pairs of estimated AAWS ($\vec{w}^i_{\text{sim}}$), (C) and the correlation between assigned AAWS ($\vec{h}$) and estimated ones ($\vec{w}^i_{\text{sim}}$). In (A–C), a random peptide library with associated AACM ($\mathbf{X}_{\text{sim}}$) of 255 14-mers and assigned AAWS ($\vec{h}$) were generated once and were kept constant across all simulation runs. For each noise amplitude, 50 simulations with a newly generated unbiased mixture were run. Antibody binding profiles were computed using Equation 4.1. Corresponding AAWS ($\vec{w}^i_{\text{sim}}$) were determined using Equation 4.8. Gaussian noise was introduced into the signal intensities: $\mathcal{N}(\mu = 0, \sigma = x)$ with x ranging from 0 to 0.1.

8.1.1 Summary I

- Increasing the amplitude of Gaussian noise introduced into simulated signal intensities (Figures 8.1 and 8.4) decreases the predictive performance (Q^2), the pairwise correlation of AAWS and recovery of assigned AAWS $\vec{h}$ (Section 8.1, Figure 8.1). Even though high noise amplitudes reduce predictive performance to near zero (Figure 8.1A), the median Pearson correlation of assigned AAWS and estimated ones only decreases to about $r = 0.5$ (Figure 8.1C).

- Across all noise levels, recovered signal intensities show higher correlation coefficients with original signal intensities than with noise-introduced ones (Section 8.1, Figure 8.2).

- The number of significantly predictive latent components is reduced with increased noise levels (Figure 8.3).

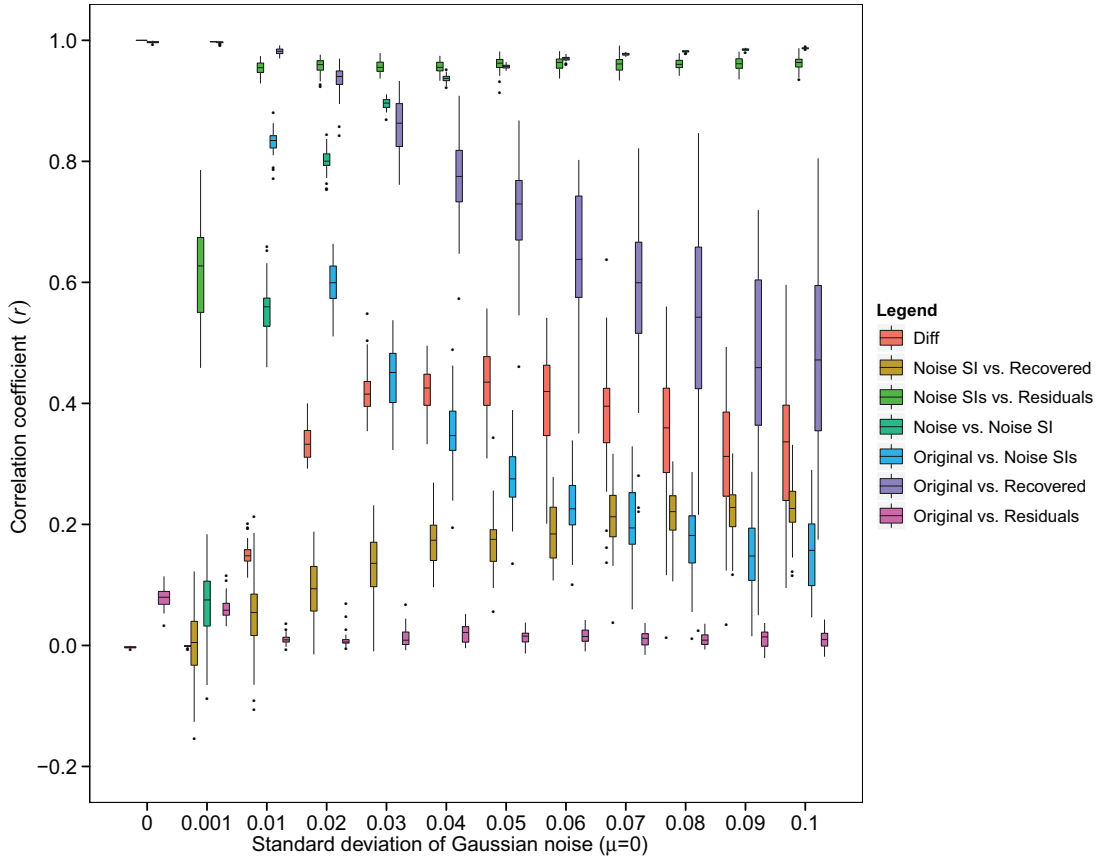


Figure 8.2: Simulations show that for unbiased mixtures, recovered signal intensities ($\hat{S}_{\text{recovered}} = \mathbf{X}\vec{w}$) (*Original vs. Recovered*) show a higher Pearson correlation with original signal intensities than with noise-altered ones (*Original vs. Noise SIs*) across all tested noise magnitudes. Signal intensity profiles are those simulated for Figure 8.1. Legend: "Diff" is the pairwise difference in correlation coefficients of "Original vs. Recovered" and "Original vs. Noise SIs" intensities; Noise: Gaussian noise; Noise SI: noise-introduced signal intensities; Recovered: recovered signal intensities; Original: original normalized signal intensities without noise; Residuals: residuals $\vec{\varepsilon}$ determined with the regression model (Equation 4.8).

8.2 The decrease in predictive performance upon introduction of noise into signal intensities depends on both peptide library size and peptide length

8.2.1 The impact of noise on the recovery of assigned AAWS is dependent on peptide library size

The predictive performance and recovery of assigned AAWS were shown above to depend only slightly on peptide library size in the case of unbiased mixtures (Section 6.1, Figure 6.1). However, when Gaussian noise is introduced into antibody binding profiles, the influence of peptide library size on recovery of assigned AAWS is increased within the

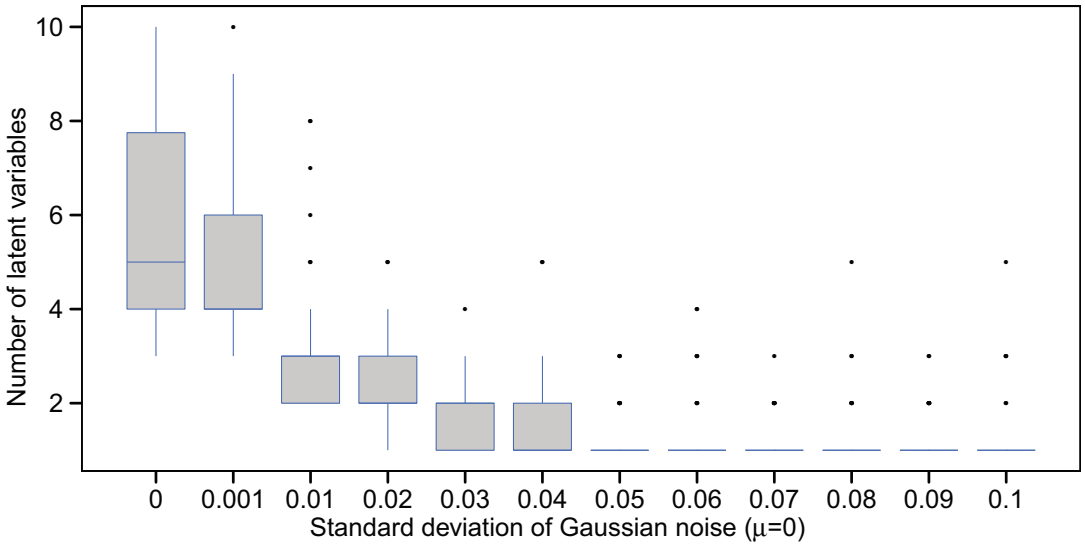


Figure 8.3: The optimal number of latent variables found with PLSR for each simulation run in Figure 8.1 decreases with increasing Gaussian noise introduced into simulated signal intensities. The optimal number of latent components was determined by evaluating the best predictive performance value (Q^2) for a given number of components (Section 3.7.2).

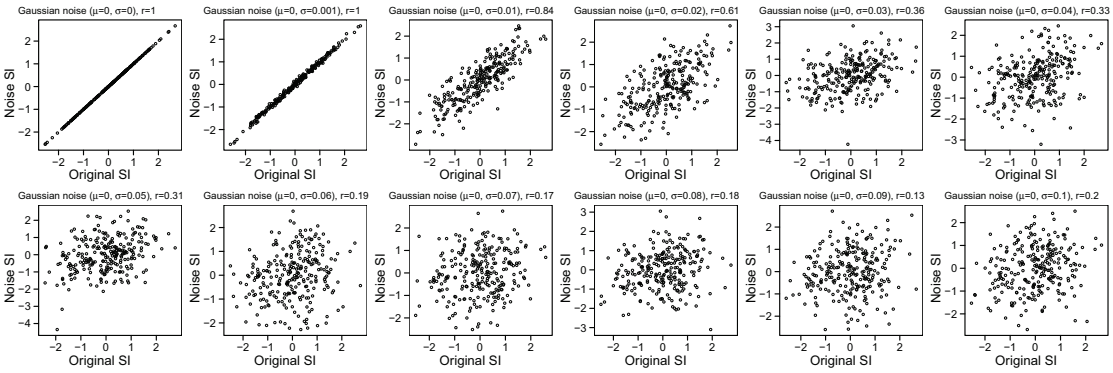


Figure 8.4: Simulated original signal intensities (*Original SI*) versus noise-introduced ones (*Noise SI*) are displayed and Pearson correlated. For each noise amplitude ($\mathcal{N}(\mu = 0, \sigma = 0 : 0.1)$), the normalized simulated signal intensities of one simulation run of Figure 8.1 are shown.

tested range (50–5000 peptides): even though the predictive performance is zero at a noise amplitude of $\mathcal{N}(\mu = 0, \sigma = 0.1)$ (Figure 8.5A), the recovery of both assigned AAWS (Figure 8.5C) and of original simulated signal intensities (Figure 8.5E) tends to perfection at high peptide library sizes (5000 peptides).

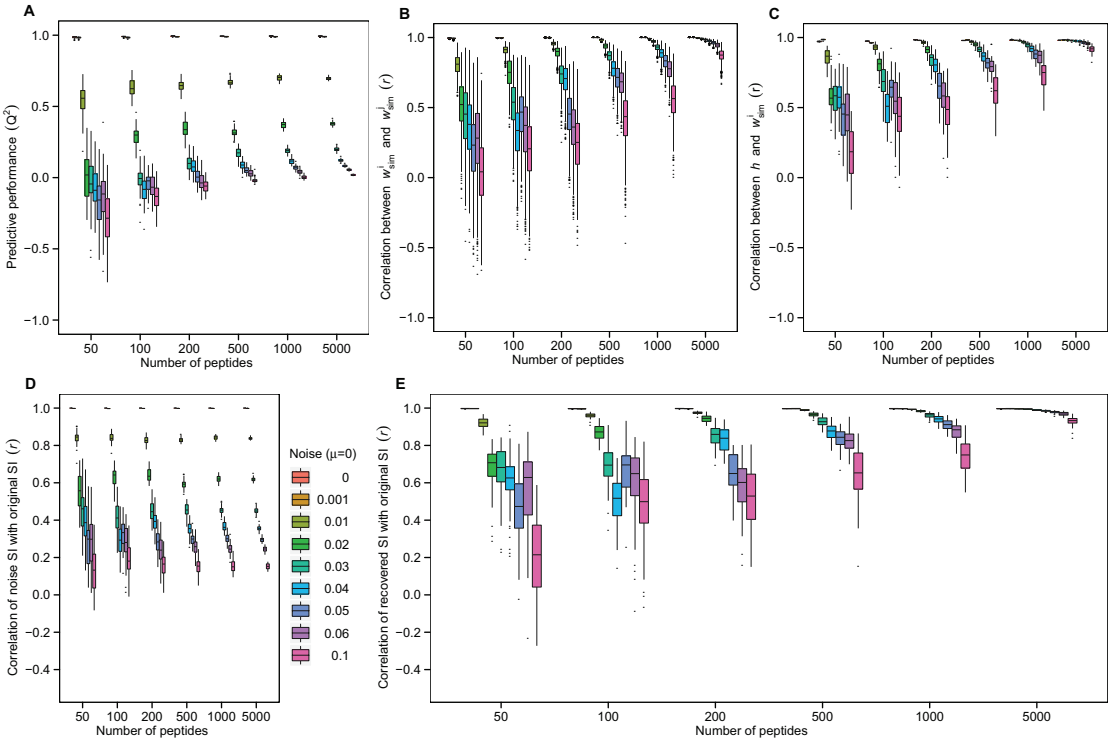


Figure 8.5: In the presence of noise, the recovery of assigned AAWS is positively dependent on peptide library size. Antibody binding profiles of unbiased mixtures ($n_{\text{Ab}} = 10000$) were simulated with varying noise amplitudes and peptide library sizes. (A) Predictive performance (Q^2), (B) Pearson correlation (r) between all pairs of estimated AAWS ($\vec{w}^i_{\text{sim}}$) as well as (C) the recovery of assigned AAWS ($\vec{h}$) are shown in function of peptide library size and noise amplitude. (D) The correlation of normalized original signal intensity profiles (signal intensities before introduction of noise, *original SI*) with noise-introduced ones (*noise SI*) and (E) the correlation of original with recovered signal intensity profiles ($\hat{S}_{\text{recovered}} = \mathbf{X}\vec{w}$, *recovered SI*) are shown. In (A–E), random peptide libraries with associated AACMs ($\mathbf{X}_{\text{sim}}$) of varying size (50–5000 peptides) with 14-mers and assigned AAWS ($\vec{h}$) were generated once and were kept constant across all simulation runs. For each peptide library size and noise amplitude, 50 simulations with newly generated unbiased mixtures were run. Antibody binding profiles were computed using Equation 4.1. Corresponding AAWS ($\vec{w}^i_{\text{sim}}$) were determined using Equation 4.8. Gaussian noise was introduced into signal intensity profiles with $\mathcal{N}(\mu = 0, \sigma = x)$ with x ranging from 0 to 0.1.

8.2.2 The impact of noise on the recovery of assigned AAWS is dependent on peptide length

The peptide length contributes to the modulation of the impact of noise on predictive performance and recovery of signal intensities and assigned AAWS. For the tested range of peptide lengths[2] (5–25 amino acids), the susceptibility to noise increases with increasing peptide length (Figures 8.6). Thus, noise increases the negative effect the peptide length

[2]The size of antibody binding sites is reported to lie within the tested range of peptide lengths (Section 1.4.2).

has on predictive performance (Q^2) and recovery of assigned AAWS $\vec{h}$ (Figure 6.1).

Simulations are consistent with the experimental data, in that for a given peptide library, 13-mers show consistently higher predictive performance values than 15-mers both for human and murine sera (Figure 5.4). Nevertheless, AAWS of 13-mer and 15-mer-libraries are highly correlated: usually Pearson correlation coefficients are above $r = 0.9$ (data not shown).

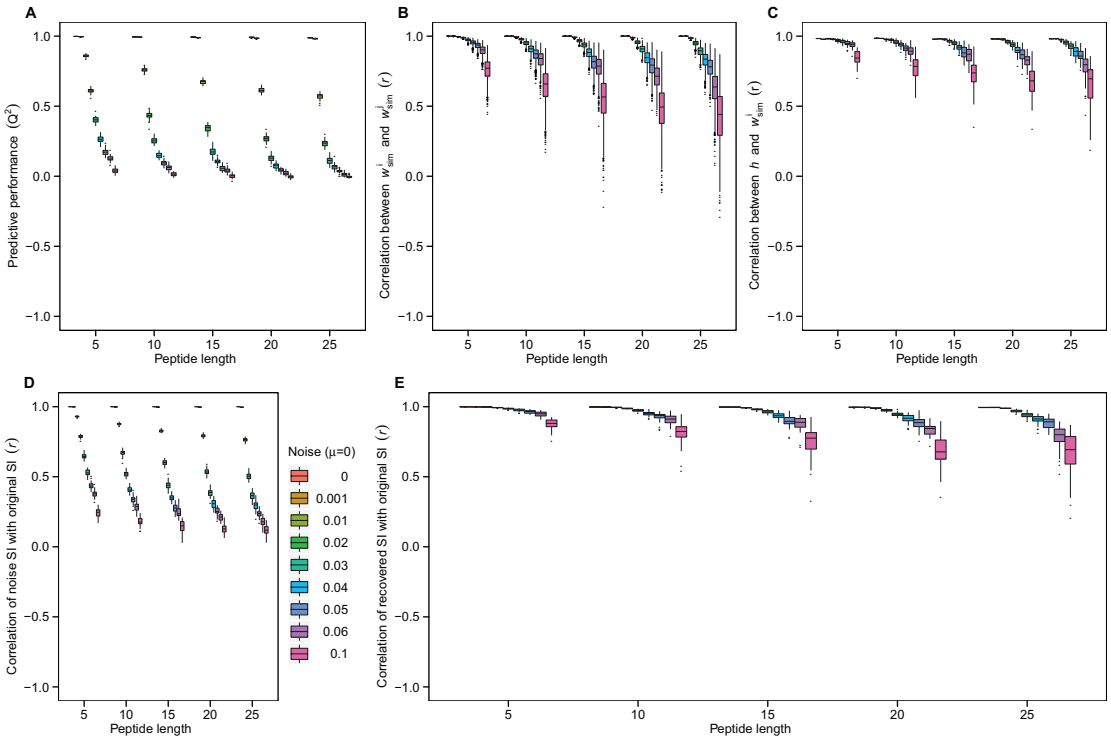


Figure 8.6: Higher peptide lengths render predictive performance, pairwise correlation and recovery of assigned signal intensities and AAWS ($\vec{h}$) more susceptible to noise. Antibody binding profiles of unbiased mixtures ($n_{\text{Ab}} = 10000$) were simulated at varying noise amplitudes and peptide lengths. (A) Predictive performance (Q^2), (B) Pearson correlation (r) between all pairs of estimated AAWS ($\vec{w}^i_{\text{sim}}$) as well as (C) the recovery of assigned AAWS ($\vec{h}$) are shown in function of peptide length and noise amplitude. (D) The correlation of normalized original signal intensity profiles (signal intensities before introduction of noise, *original SI*) with noise-introduced ones (*noise SI*) is shown as is (E) the correlation of original with recovered signal intensity profiles ($\hat{S}_{\text{recovered}} = \mathbf{X}\vec{w}$, *recovered SI*). In (A–E), random peptide libraries with associated AACMs ($\mathbf{X}_{\text{sim}}$) of 1000 14-mers and assigned AAWS ($\vec{h}$) were generated once and were kept constant across all simulation runs. Peptide lengths were varied from 5 to 25 amino acids. For each peptide length and noise amplitude, 50 simulations with newly generated unbiased mixtures were run. Antibody binding profiles were computed using Equation 4.1. Corresponding AAWS ($\vec{w}^i_{\text{sim}}$) were determined using Equation 4.8. Gaussian noise was introduced multiplicatively into signal intensity profiles with ($\mathcal{N}(\mu = 0, \sigma = x)$ with x ranging from 0 to 0.1.

8.2.3 Summary II

- In the presence of noise, the recovery of assigned AAWS ($\vec{h}$) (Section 8.2.1, Figure 8.5C) and original simulated signal intensities (Figure 8.5E) depends positively on peptide library size.
- Higher peptide lengths render predictive performance (Q^2, Figure 8.6A), recovery of assigned AAWS ($\vec{h}$) (Section 8.2.2, Figure 8.6C) as well as the recovery of original signal intensities increasingly susceptible to noise (Figure 8.6E).

8.3 Estimated AAWS are consistent across microarray batches but differ by manufacturer and species

AAWS of healthy[3] individuals of experimental studies described in *Methods* (Section 3.5, $n = 158$) were correlated and presented in a heatmap to study their Pearson correlation-based clustering. For inclusion in the heatmap, AAWS had to originate from the following libraries: $J^{255}_{\text{14-mer}}$, $J^{3352}_{\text{15-mer}}$, $J^{3418}_{\text{15-mer}}$, $J^{3626}_{\text{15-mer}}$ or $P^{942}_{\text{15-mer}}$. AAWS based on 13-mer libraries were excluded.

The heatmap (Figure 8.7) shows clustering of AAWS by species and manufacturer. Exemplary representative comparisons of AAWS between pairs of experimental studies are summarized in Table 8.1 and Section A.6.

8.3.1 Summary III

- The correlation of AAWS across experimental studies is high provided that (i) AAWS are compared among studies performed on the same platform (JPT, Pepscan) and (ii) the compared samples originate from the same species (human, mouse) (Table 8.1, Appendix A.5, Figures S.5–S.11). Indeed, AAWS cluster by species (mouse, human) and by manufacturer (JPT, Pepscan, Figure 8.7). AAWS clustering by species and manufacturer is independent of the correlation method (Pearson, Spearman) used (Figure S.12).
- AAWS are consistent across (i) murine genetic backgrounds (NOD, C57BL/6, BALB/c) and (ii) microarray batches (Section A.6, Figure 8.7, Table 8.1). The batch consistency of AAWS contrasts with the batch inconsistency of signal intensity profiles (Section 3.5, Figure S.9, Table S.2, [259]).

[3] Analogous to Figure 5.4 in Chapter 5, for both the NephroFIT and the NephroFIT-Pepscan study, "No rejection" individuals were assessed in Figure 8.7 because no more than two samples of healthy controls were incubated in these studies.

Manufacturer, Species	Pairwise comparison of studies by AAWS	Overall median correlation coefficient	Reference
JPT, Human	SHS, Glioma 09	$r = 0.73$	Figure 8.7
	Glioma 09, NephroFIT	$r = 0.74$	Figure S.8
	Glioma 08, Glioma 09	$r = 0.77$	Table S.2
JPT, Mouse	MS (BALB/c, C57BL/6)	$r = 0.91$	Figure S.5
	NS (C57BL/6), MS (C57BL/6)	$r = 0.77$	Figure S.6
JPT, Human versus Mouse	MS (BALB/c), SHS	$r = -0.28$	Figure S.7
	NS (C57BL/6), SHS	$r = 0.26$	Figure 8.7
Pepscan, Human	NephrOT, NephroFIT-Pepscan	$r = 0.76$	Figure S.9
Pepscan versus JPT, Human	SHS, NephrOT	$r = -0.02$	Figure 8.7
	NephroFIT, NephroFIT-Pepscan	$r = 0.10$	Figure S.10
Pepscan versus JPT, Mouse versus Human	MS (BALB/c), NephrOT	$r = -0.02$	Figure S.11

Table 8.1: The correlation of AAWS between experimental studies is high provided that (i) AAWS are compared among studies performed on the same platform (JPT, Pepscan) and (ii) the compared samples originate from the same species (human, mouse). Exemplarily, but representatively, AAWS of healthy individuals used in Figure 8.7 were Pearson correlated between experimental studies in a pairwise fashion. The term *"overall median correlation coefficient"* denotes the median of all unique Pearson correlation coefficients for a comparison of AAWS of two experimental studies. Please refer to Appendix A.5 for the corresponding graphs. Legend: MS; Mouse study, NS; NOD study, SHS; Slovenian healthy study.

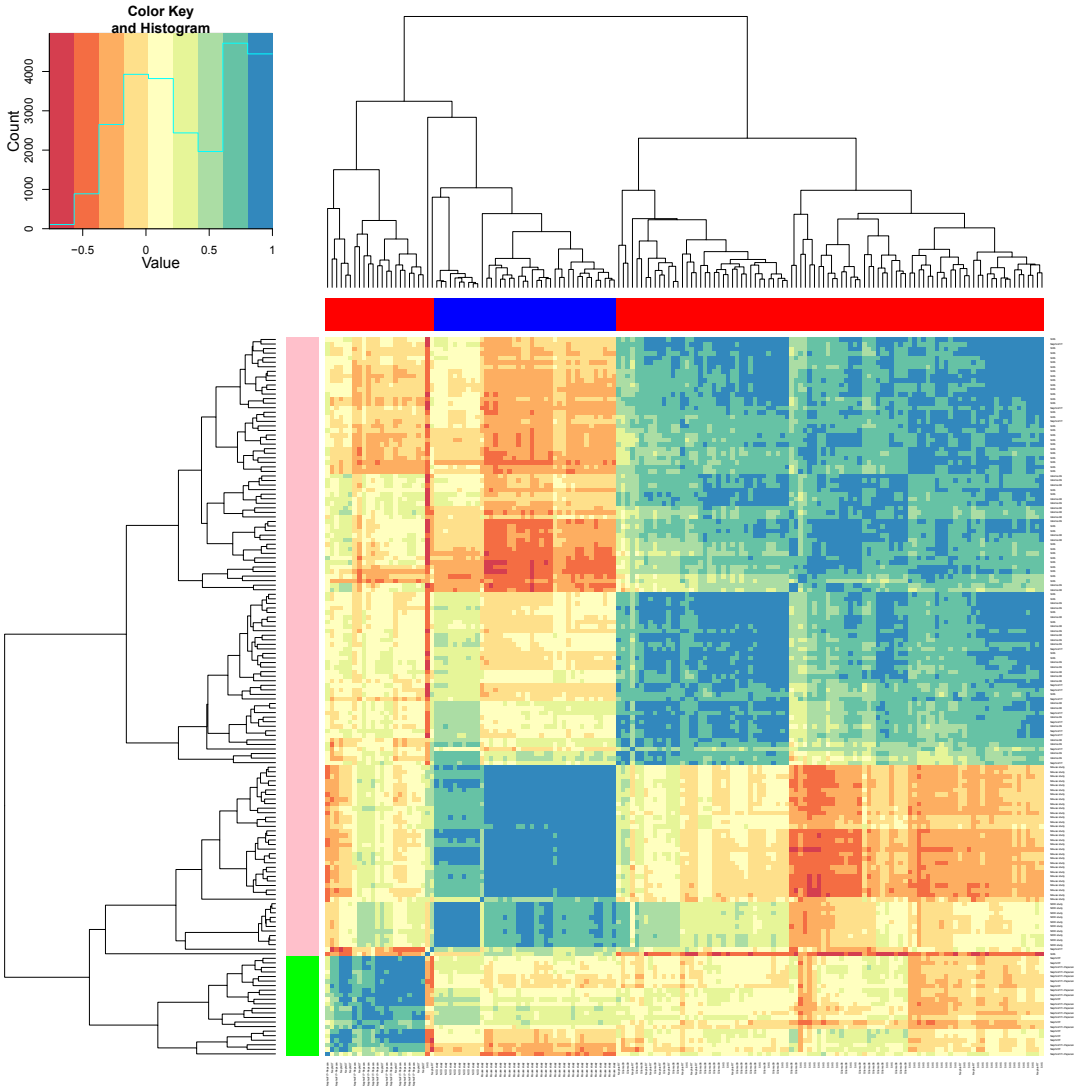


Figure 8.7: Heatmap of AAWS of healthy individuals of all experimental studies assessed in this thesis shows Pearson-based clustering of AAWS by species (human: red, mouse: blue) and manufacturer (JPT: pink, Pepscan: green). Number of total AAWS clustered: 158. (For both the NephroFIT and the NephroFIT-Pepscan study, "No rejection" individuals were assessed in this Figure, because no more than two samples of healthy controls were incubated in these studies). AAWS of 13-mer libraries were excluded from this thesis. The weight of cysteine (C) was removed from all AAWS of cysteine-containing libraries ($J^{255}_{14\text{-mer}}$, Table 3.1). Antibody binding profiles were measured with the respective libraries and AAWS were determined with Equation 4.8. The heatmap was built by Pearson-based correlation as detailed in Section 3.9.2. The correlation method (Pearson, Spearman) has no major impact on the clustering by species or manufacturer (Figure S.12).

9 Discussion

Parts of this Chapter were recently published [104].

9.1 Assessing the consistency of in silico and in vitro antibody-peptide reactivity data

9.1.1 Unbiased antibody mixtures show ensemble properties

A mathematical model for antibody-peptide binding based on the law of mass action was proposed (Section 4.2). Herein, the binding signals for a given simulated monoclonal antibody depend nonlinearly on the amino acid's position in a given peptide. In this context, a property vector $\vec{h}$ was defined, which characterizes each peptide's amino acid binding strength. The model was analyzed and ported to a simulation framework.

The analysis of the model highlighted ensemble properties of a special case of mixtures, termed unbiased, which were defined as having (i) a highly diverse, (ii) i.i.d. antibody repertoire, and (iii) showing no antibody dominance. Antibody dominance is defined as the relative concentration increase of a small portion of the antibody mixture (Section 4.3). The derived ensemble properties of unbiased mixtures result in a *linear* relationship of peptide signal intensity and peptide sequence. This linear relationship led to the formulation of a linearly solvable regression model. The regression model yields estimated AAWS ($\vec{w}$) as near perfect predictors of signal intensity profiles, which recover assigned AAWS ($\vec{h}$). For antibody-dominated—also termed biased—mixtures, the predictive performance of the regression model depends on the bias (Section 4.3.4, Figure 4.3).

The ensemble properties' implications for signal intensity profiles of unbiased mixtures are twofold: (i) the simulated signal intensity profile is specific for a given peptide library with given assigned AAWS: it is obtained independently of any unbiased mixture (Figure 4.2). (ii) AAWS are a compact, lossless representation of antibody binding profiles. They represent a (near perfect) dimension reduction of antibody binding profiles from the signal intensity space onto the amino acid space (Section A.8.2). In this amino acid space, the peptide's amino acid composition, and no longer its ordered sequence, is determinative of a peptide's signal intensity (Sections 4.3.2 and A.8.2).

In this work, antibody sequences were drawn from a *Gaussian* distribution in order to analytically derive the exclusive dependence of signal intensities on peptide amino acid composition for unbiased mixtures (Equation 4.7). Nonetheless, all simulations, being performed with an underlying *uniform* distribution, indicated that mathematical predictions and simulation results were in full accord (Section 4.3). The choice of the distribution is rather outweighed by the need for small components both in assigned AAWS ($\vec{h}$) and antibody sequences (Footnote 3, page 53).

The regression model's predictive performance was found to be robust against changes (i) in the total antibody concentration (Sections 6.2 and 6.3), (ii) in the distribution of assigned AAWS (Section 6.4) and (iii) in the simulated peptide library (Sections 4.3.3 and 6.1). However, the predictive performance is sensitive to a violation of the i.i.d. generation of antibody repertoires (Figure 6.11), a biasing of the repertoire (antibody dominance, Figure 4.3) and an introduction of noise into simulated signal intensities (Section 8.1).

9.1.2 Predictions of the mathematical model are validated by in vitro antibody-peptide reactivity data

The major prediction of the proposed mathematical model is that highly diverse serum antibody mixtures—in contrast to low-diversity antibody mixtures—show a convergence to a signal intensity profile which is linearly predictable to near perfection by the peptides' amino acid composition (AACM). In this thesis, two extreme cases were used to validate this prediction. Monoclonal and serum antibodies[1] were incubated on the same peptide library: the predictive performance, a measure for the quality of the AACM-based regression model, was significantly higher for serum than for monoclonal antibodies (Figure 5.1).

In accord with the mathematical predictions, AAWS, and to a lesser extent signal intensity profiles, were largely consistent across healthy individuals, both human and murine (Figure 5.4), and therefore independent of an individual's serum antibody composition. While it could be argued that antibody mixtures of inbred mice as used in the Mouse study (Section 3.5.8) are likely to be similar[2], this argument does certainly not hold with respect to the utilized human sera. Thus, serum antibodies show ensemble properties that were predicted by the mathematical model.

A corollary of the derivation of serum antibody ensemble properties (Equation 4.6) is that antibody dominance leads to a decrease in predictive performance (Sections 4.3.4 and 6.5), Figure 4.3). It is known that during a primary immune response in immunocompetent hosts, antigen-specific antibodies are produced in high abundance [2, 256]. Therefore, it could be expected that sera of infected mice show reduced predictive performance values, altered measured signal intensity profiles and estimated AAWS, as was confirmed by experimental data (Figures 5.2, S.2 and S.3).

[1] *Human* monoclonal and *murine* serum antibodies were incubated to validate the mathematical predictions (Figure 5.1). The antibodies' origin is unimportant as both of these entities are studied separately and only predictive performance values (Q^2) are being compared. Murine serum *IgG* antibodies yield predictive performance values similar to those of the shown IgM isotype [104].

[2] Mice bred under SPF conditions (Section 3.5.8) were also shown to harbor GCs even in non-immunized conditions [123]. This suggests a mouse-to-mouse variation of antibody repertoires. Also, AAWS were similar across mouse strains (Figures 8.7 and S.5).

9.1.3 The concept of unbiased mixtures best approximates sera of healthy individuals

In view of the relatively high predictive performance of antibody binding profiles of serum samples of healthy individuals, it was postulated that these sera exhibit properties of unbiased mixtures [104].

The first prerequisite for an unbiased mixture is high diversity (Section 4.3.3). This requirement appears to be met since at any given time at which an individual is not undergoing an acute immune response, the composition of serum antibodies is reported to be provided mostly by a rather constant number of antibody secreting cells (Section 1.6, [112, 290]). The functional IgM diversity is estimated to be of the order of 10^4 clones [103][3] and the potential antibody diversity is very high [79].

However, the fulfillment of the second requirement, the independent and identical distribution of antibody binding sites, is harder to claim. Even though the antibody repertoire is composed of germ line sequences and shaped by clonal selection, V(D)J recombination and—at later stages of the immune response—somatic hypermutation arrange and mutate these segments in a largely random fashion [3]. This is consistent with the hypothesis that antibody repertoires can potentially recognize the entire antigenic universe (Section 1.4.4, [66, 291]).

The third requirement, the absence of antibody dominance, is likely not met by sera of healthy individuals since this would entail a complete lack of specificity[4]. Indeed, data from the Slovenian healthy study showed evidence of slight antibody dominance which I suggested to be responsible for the clustering of ranks of antibody profiles by healthy volunteer (Sections 6.3 and 9.2.1). The presence of antibody dominance and correlated antibody repertoires likely also contributed to the fact that the predictive performance values did not reach perfection in vitro (Figure 5.4) as it was the case for unbiased mixtures (Figure 4.2A). In addition, low predictive performance values may also be due to noise (Section 9.2.3).

9.2 Discussion of results in light of antibody profiling and serological diagnostics

The experimental validation of the mathematical model's predictions unifies the simulation framework and the experimental system used into a novel integrated systems framework

[3] The correspondence of the diversity of the BCR repertoire (Section 1.5.3) and that of the antibody repertoire remains as of yet unclear (Section 1.6.3).

[4] Unbiased mixtures are amino acid composition-specific (Sections 4.3.1 and 9.1.1). Thus, two peptide sequences, which differ by amino acid order but not composition, could not be differentiated by an unbiased mixture. Therefore, if sera of healthy individuals were to function as unbiased mixtures, the fine differentiation between sequences would be abolished to a great extent. In fact, Bachmann and colleagues showed that in vivo protective capacities of a panel of monoclonal IgG antibodies to the VSV glycoprotein were independent of Ig subclass, avidity, neutralization rate constant, and in vitro neutralizing activity; above a minimal avidity threshold, protection depended simply on a minimum serum concentration [292]—or in terms of this thesis—on a minimum level of antibody dominance.

for both the theoretico-statistical and experimental analysis of antibody-peptide reactivity data [293].

9.2.1 Assessing the classifiability of antibody binding profiles

The analysis of the mathematical model suggests that differences in antibody binding profiles can only be realized by (i) changes in the total antibody concentration (Section 6.3) or (ii) by relative antibody concentration differences (antibody dominance, correlated antibody repertoires, Sections 4.3.4, 6.5 and 7.1).

Both types of concentration changes were shown to have an impact on the clustering of simulated signal intensity profiles, depending on the correlation coefficient (Pearson, Spearman) used (Section 6.2, Figures 4.3, 6.9 and S.13, Table S.3). Due to the monotonic influence of total antibody concentration on simulated signal intensities (Figure 6.4A), total concentration differences have no bearing on their ranks (Figure 6.4B), Only a relative change in the antibody composition of the tested mixture, which is caused by antibody dominance, can induce a change of ranks (Figure 4.3).

In summary, in silico findings suggest the following for serological diagnostics: if total concentration varies across antibody mixtures, both antibody dominance and total antibody concentration differences can cause group-specific antibody binding profiles, thus allowing for a separation of antibody mixtures. If the total antibody concentration is constant across antibody mixtures, only a *shared* antibody dominance allows a separation of antibody binding profiles by inducing a higher inter-group than intra-group variance (Figure 7.1C): randomly biased antibody mixtures lead to *isotropically* distributed signal intensity profiles in the variance space (Figure 7.1A).

The intra-group consistency of antibody-peptide binding behavior is a basic premise of serological diagnostics (Section 1.8): the mathematical model does not only fulfill this premise, but also predicts antibody dominance[5] as a condition which is able to establish classifiable intra-group consistency.

Thus, the fact that signal intensity profiles in the Slovenian healthy study (SHS) show, in addition to Pearson-based clustering (Figure 6.8A and B), also Spearman-based clustering by healthy volunteer (Figure 6.8C and D), suggests that (i) most tested human volunteers possess dominant antibodies with different peptide-library binding behavior, (ii) which did not change over the course of the experiment. Individual antibody dominance would thus prevail in its influence on clustering over the evidenced total antibody concentration differences (Figure 6.7).

Baseline studies such as the SHS are important for serological diagnostics as they show that even if antibody binding profiles and AAWS are relatively highly correlated due to suggested ensemble properties of serum antibodies (Figure 5.4), sera are still biased enough to exhibit clustering. Consequently, if disease groups are to be separated based on their profiles, disease-induced dominant antibodies must have a stronger influence on binding signals than the individual bias. The individual variability must be taken into account for serological diagnostics to detect group-specific differences in antibody

[5] Unbiased mixtures are special cases of biased ones (Section 4.3.3).

binding profiles.

Samples of BALB/c mice (Section 3.5.8, Figure 3.1) were divided into three groups : *healthy*, *acute phase* and *early chronic* phase. These groups could be differentiated by PCA based on signal intensity[6], ranks[7] and AAWS (Figures 5.3, S.2 and S.3). Both, signal intensities and ranks, showed high classification accuracies using only a small percentage of the analyzed library (Section 5.3, Table 5.3). The high balanced accuracy was accompanied by a reduction of (i) the predictive performance (Figure 5.2A), (ii) the pairwise correlation of signal intensity profiles (Table 5.1) and (iii) AAWS (Figure 5.2B). The above described simulations results thus suggest that a shared (consistent) antibody dominance contributed primarily to the classification of the murine antibody binding profiles. The dependence of mean signal intensity on antibody concentration was low and not significant (Table 6.1). Rank changes during acute immune responses were also noted by Stafford and colleagues[8] [210].

Convergent—or shared—B-cell responses against an antigen across individuals represent a phenomenon that has long been described. For some antigens, this has been characterized as repertoire bias, where particular genes from the germline repertoire are favored in the panel of antibodies that is raised during the immune process [295–297]. In addition to repertoire bias, identical CDR3 sequences were found to be shared across rodents immunized with both Tetanus toxin [298] and other peptide antigens [295]. In addition, matching CDR3 sequences were also observed across healthy individuals (human, zebra fish) [93, 299].

It is unknown, however, how in vitro antibody dominance relates to in silico antibody dominance (see also Section 9.1.2). Simulated mixtures start to show ensemble properties if their number exceeds 50 antibodies (Figures 4.2 and 4.3). The human Tetanus-specific repertoire has been estimated to be of the order of 100 distinct B-cell clones [298, 300]. Similarly, the murine VSV-specific repertoire shows about 10^2 to 10^4 specificities [135]. Further, humoral immune responses were found to be correlated [298] such that—according to simulation results—also higher clonal diversities could be responsible for a decrease in predictive performance (Figure 6.11).

9.2.2 The profiles of unbiased mixtures are crucial to isolating the signal of dominant antibodies

If dominant antibodies are important for the classification of sera (Section 9.2.1), it would be of interest to isolate their signal from the background [290]. Assuming the background to behave as an unbiased mixture, the signal of the dominant antibody can be, to a

[6] In light of the high classificatory power of IgM profiles, it is worth noting that IgM has been shown to be *non*-protective against HB (Section 1.8.2, [256]).

[7] While ranks represent a highly helpful means to control for the origin of high classificatory power of a data set (total antibody concentration or antibody dominance), such as in the early stages of lead discovery, their usefulness remains limited for multi-parameter diagnostics which pursues the goal to single out a small subset of peptides with high classificatory power (see e.g. [210, 294]): the classificatory power of ranks decreases with decreasing peptide library size.

[8] "Sera from infected individuals demonstrate generally higher reactivity for some peptides, but some peptides indicate less reactivity relative to normal controls" (Figure 8 in Stafford and colleagues [210]).

certain extent, isolated (Section 4.3.5, Equation 4.9). Since assigned AAWS are likely to be recovered well by sera from healthy individuals (Section 9.1.3), they could be used to recover the signal of an unbiased serum ($\vec{S} = \mathbf{X}\vec{w}_{\text{healthy serum}}$), which can subsequently be used for the isolation of the signal of dominant antibodies.

To test this mathematical concept in vitro, further experiments involving incubations of monoclonal antibodies of different concentration mixed with serum from healthy individuals have to be performed. Such experiments have been recently published by Stafford and colleagues who show that the signal of antibodies diluted in 10x and 100x excess IgG (biased mixture) still allowed high correlations of the signal of the monoclonal antibody with the biased antibody mixture [210]. If the derived formula (Equation 4.9) held true in vitro, the correlation of the dilution of the monoclonal antibody and the recovery of its signal would have to be determined. In fact, simulations showed that this correlation depends both on the dilution and the diversity of the dominant antibodies (Figure 4.4). Replicating simulations in vitro would allow the determination of a threshold giving the minimum concentration necessary for a faithful isolation of the dominant antibody's signal.

9.2.3 Technological implications of the AAWS concept

Amino acid-associated weights are a means for comparing the signal of peptide libraries across microarray batches

The methodological variability of antibody profiling studies is high (Section 1.8.1, Table 1.1, [245]). Studies vary with respect to peptide manufacturer, peptide library (random-sequence peptides, protein libraries), dilution of sera (dilution of serum/plasma, normalization of Ig-concentration across samples) as well as data normalization and bioinformatic means used to analyze antibody-peptide reactivity data. Adding to the methodological variability of studies is the batch-to-batch variability of (random) peptide arrays [259]. In fact, peptide microarray studies are far less comparable [238, 245] than cDNA based microarray studies, which are performed based on highly standardized protocols [240–244].

AAWS of healthy individuals were found to be highly consistent by study (Section 5.4). This is in accord with the mathematical prediction that for a given peptide library AAWS are not function of the composition variability of unbiased mixtures (Sections 4.3.1 and 9.1.2). Due to the minimality of the proposed mathematical model, the influence of parameters such as species, microarray batch and manufacturer on the estimation of AAWS is, however, unknown. In the following, the parameters' impact on AAWS is discussed.

AAWS of human as well as murine samples incubated on the JPT platform were found to be consistent *across* batches and peptide libraries (Section A.6, Figure 8.7, Table 8.1), which contrasts with a high *signal intensity* batch variability (Section 3.5, Figure S.9, Table S.2, [259]). In addition, murine AAWS are relatively consistent across genetic backgrounds (NOD, BALB/c, C57BL/6, Figures 8.7, S.5 and S.6). Likewise, AAWS determined from human sera incubated with Pepscan arrays showed considerable

consistency across batches (Figures 8.7 and S.9).

In contrast, Pepscan-AAWS correlated poorly with both human- and mouse-AAWS determined on the JPT-platform (Figures 8.7 and S.12, Table 8.1). Pepscan-AAWS showed highest correlations with tested physico-chemical properties (Table 5.4) and propensity scales (Table 5.5). Thus, with respect to the studied batches in this thesis, the estimation of AAWS is influenced by the array-producing company. In fact, production of arrays and peptides[9] differed by manufacturer (Section 3.1.1), which could explain the observed difference in estimated AAWS.

The influence of the printing technology on estimated AAWS seems limited as C57BL/6-AAWS of the NOD study (JPT-contact) were highly correlated to C57BL/6-AAWS of the Mouse study (JPT-non-contact) despite the high difference in predictive performance values[10] (Figures 5.4 and S.6). However, the non-contact-printed $J^{255}_{\text{14-mer}}$ library was the only one not showing a significant correlation with total IgM concentration (Table 6.1) hinting to the fact that printing methods influence antibody-peptide binding to a certain extent.

Furthermore, AAWS of human and murine samples (on the JPT-platform) correlated poorly; while C57BL/6-AAWS of the NOD study were relatively highly correlated to C57BL/6-AAWS of the Mouse study (Figure S.6), they only showed low correlations with AAWS of human serum/plasma samples. How could these differences arise? According to the minimal model (Section 4.2), these differences can either result from different antibody dominances or different (printing-specific) assigned AAWS ($\vec{h}$). (i) However, since murine AAWS both on non-contact and contact-printed arrays where highly correlated (Figure S.6), different assigned AAWS are unlikely the reason for the found discrepancy. (ii) The poor correlation of murine and human AAWS cannot be explained by the low predictive performance[11] of human samples on 15-mer JPT-libraries since 13-mer-AAWS—which show increased predictive performance values—and 15-mer-AAWS are highly correlated (Section 8.2.2). (iii) AAWS of human and murine sera differed in their correlation pattern with estimated AAWS of human monoclonal antibodies (Figure 7.4). Thus, differences in human and murine AAWS could either result from different antibody repertoires or relative antibody concentration differences in antibody mixtures (different antibody dominances). Indeed, murine and human antibody repertoires are quite different from one another (Section 1.5.1, [84]). In order to further study the dependence of AAWS on species, incubations of peptide arrays of the same batch with human and murine serum samples are needed.

Taken together, it has been shown that the manufacturing process (JPT versus Pepscan, Section 3.1) has a considerable impact on estimated AAWS. This difference in estimated AAWS could be due to differences in assigned AAWS which would have—according to the model—consequences for the binding of antibody mixtures to peptides. Manufacturer-dependent assigned AAWS imply that the same peptide library, made by different manufacturers, is not only likely to yield absolute but also relative signal intensity

[9] The majority of peptides are likely not folded or are only transiently non-linear [210].

[10] Differences in predictive performance values were high if comparing 14- to 15-mers, but rather low for 13-mers compared to 14-mers (Figure 5.4).

[11] A low predictive performance could lead to a biased estimation of AAWS (Figure 4.2).

differences: peptides would show different ranks in function of the difference in assigned AAWS[12]. It is therefore possible that with respect to one peptide library, serum samples clustered by disease status whereas the other one would fail to show a discrimination of samples. Indeed, simulations indicated that changing the assigned AAWS ($\vec{h}$) of a peptide library altered monoclonal antibody binding behavior: strong antibodies turned into weak ones (Section 7.3, Figure 7.3).

Thus, AAWS could be useful for array manufacturing companies for assessing the batch-to-batch bias with respect to the underlying amino acid binding behavior. This could increase rank consistency and thereby comparability of measured signals across microarray batches. Furthermore, the AAWS may be a beneficial concept for normalization approaches[13]. Indeed, AAWS have, in contrast to signal intensity profiles, the advantage of being independent of the peptide library's dimension (Figures 8.7 and S.12). In this thesis, numerous types of peptide libraries were used (Table 3.1), which are impossible to compare based on their signal intensity profiles.

Recovery of assigned AAWS is high due to the noise resistance of estimated AAWS

Low predictive performance values, may—in addition to being caused by the violation of the assumptions of unbiased mixtures—be a result of technological noise. Noise could originate from varying peptide spot quality on microarrays and by the experimental procedure itself. In this vein, it has been recently shown that peptide density has a considerable effect on measured peptide signal intensity [210, 302]. To model to some degree the influence of noise[14] on the predictive performance of profiles of unbiased mixtures, simulated signal intensity profiles were overlaid with increasing amplitudes of noise. Noise lead to a decrease in both predictive performance and recovery of assigned AAWS $\vec{h}$. However, even at higher noise amplitudes with predictive performance nearing zero, the recovery of assigned AAWS (Figure 8.1) and consequently that of signal intensities (Figure 8.2) was rather high (Section 8.1).

Simulations showed that a higher consistency of AAWS estimation is achieved with short peptides and higher peptide library sizes (Figures 8.5 and 8.6). While shorter

[12]The NephroFIT (Section 3.5.4) and the NephroFIT-Pepscan study (Section 3.5.5) are inappropriate for a comparison of balanced accuracies since these studies were incubated by (i) different researchers (Juliane Lück (NephroFIT-Pepscan study), Bodo Steckel (NephroFIT study)), (ii) the BK virus status, which was not checked for each sample beforehand, could skew results [301] and (iii) samples were incubated with different techniques (manual (NephroFIT-Pepscan study) and automated incubation (NephroFIT study)).

[13]In this work, normalization of BALB/c signal intensity profiles prior to PCA or P-SVM analysis (Chapter 5) was not performed (Section 3.4.1) due to their low technological variability compared to their high biological variability (Tables 5.1 and 5.2). Low technological variability is a phenomenon observed throughout all experimental studies (Section 3.5, page 32). In addition, this work suggests that the understanding of antibody-peptide reactivity data is incomplete. However, the application of normalization approaches must be knowledge-driven.

[14]Increasing noise amplitudes decrease the AACM-explicability of simulated signal intensity profiles of unbiased mixtures. The introduction of noise into antibody binding profiles reflects to some extent the presence of systematic noise in an in vitro setting. More generally, (i) this noise models the non-correspondence of mathematical model and *simulated* signal intensity profiles, (ii) and consequently the lack of explicative power of the here formulated model to fully reconstruct *in vitro* measurements.

peptides alleviate the problem of the approximation of linearized signal intensity[15], high peptide library sizes increase the resolution of the data[16]. If further validated in vitro, large peptide libraries with short peptides could be used by manufacturers to estimate AAWS with greater precision (Section 9.2.3).

In *simulations* the observed noise resistance of estimated AAWS is mostly owed to partial least squares regression (PLSR) which reduces the dimensionality of the data by filtering for signal variance (Figure 8.3) [270]. Thus, even though the high consistency of *in vitro* estimated AAWS across samples, microarray batches and libraries (Figure 5.4 and 8.7, Table 8.1), may ultimately be due to ensemble properties of serum antibodies, simulations suggest (Figures 8.2 and 8.3) that PLSR is driving to a large extent the consistency's detectability.

Technological features may bias amino acid-associated weights

Average AAWS differed mostly by species (mouse, human) and array platform (JPT, Pepscan) (Figure 5.5). Whereas for murine samples, mainly aromatic amino acids are top positive contributors to signal intensity, top ranking amino acids for human samples incubated on JPT arrays are methionine and proline. Average AAWS of human samples incubated on Pepscan arrays showed mainly lysine, histidine, proline and tryptophane as positively contributing amino acids.

However, average AAWS (Figure 5.5) should be interpreted with caution. In addition to being interpreted as both amino acid antibody binding preferences (Section 4.2, Table 5.5) and physico-chemical properties (Section 5.5, Table 5.4), signal intensity profiles may also be influenced by at least two other factors: (i) the accessibility of peptides and (ii) a possible interaction of aromatic amino acids and aromatic labeling dyes.

Accessibility may bias the resulting signal intensities systematically. As to the $J^{255}_{14\text{-mer}}$ library, it was found that cysteine contributes negatively to the signal intensity. This could be partly due to its ability to form disulfide bonds, which causes increased aggregation of cystein-containing peptides, and diminishes their surface exposure. Consequently, this would lead to reduced antibody-peptide binding and accordingly to a reduced signal intensity. In general, it cannot be ruled out that aromatic amino acids interact via π-stacking with the aromatic labeling dyes Alexa Fluor 546 and 647 which are coupled to the secondary antibodies. Indeed, it has recently been found that TAMRA, another aromatic dye, cross-reacts with individual amino acids in a peptide sequence [303].

In order to minimize this effect, secondary-antibody correction on the log-transformed signal intensities was performed for AAWS determination (Figure S.4).

[15] Higher peptide lengths increase the uncertainty in the model. Recall that for unbiased mixtures, the peptide signal intensity is neared by a function of the sum of components of the peptide vector (Equation 4.8). Increasing the number of terms in that sum, increases the uncertainty in the derived approximation.

[16] Higher peptide library sizes increase predictive performance as they provide the regression model with more input thereby increasing the reliability of estimated AAWS.

9.3 Discussion of results in light of B-cell epitope mapping

The prediction of linear B-cell epitopes was first done by using propensity scales (Section 1.7.3, [177, 182, 192]). These scales assign a propensity value to every amino acid based on a priori studies of their physico-chemical properties. Average AAWS (Figure 5.5) correlated poorly with widely used propensity scales for epitope prediction with the exception of Pepscan-AAWS (Table 5.5).

Blythe and Flower tested 484 amino acid propensity scales on a set of 50 epitope-mapped proteins. They found that even the best set of scales performed only marginally better than random[17] [172]. This thesis shows that unbiased mixtures represent a special case for which the converse holds true: antibody binding profiles of unbiased mixtures can be predicted based on AAWS. Further, this work suggests that the use of amino acid scales becomes increasingly less justified with increasing dominance of antibodies in a serum (Figures 4.3, 5.1 and 5.2). In fact, each of Blythe and Flower's experiments used polyclonal antibodies raised against the whole protein [172]. It can therefore be conjectured that the used polyclonal antibody mixtures were biased, meaning that they contained dominant antibodies [174]. In this regard, this work provides a possible explanation to Blythe and Flower's findings. More generally, it is suggested here that results obtained with polyclonal antibody mixtures tend to be skewed by their inherent ensemble properties, which obscure the affinities of epitope-specific antibodies.

It was suggested that (i) "overall poor performance [of B-cell epitope prediction approaches] may reflect the generality of antigenicity and hence the inability to decipher B-cell epitopes as an intrinsic feature of the protein. It is an open question as to whether ultimately discriminatory features can be found" (Section 1.4.1, [30]). (ii) Stretches of protein sequences may not per se constitute epitopes—the context of antibody-peptide binding may be crucial for epitope definition [15, 50, 174, 304]. However, even though context-related epitope definition may play a role in vivo, the fact that antibody binding profiles, at least for some diseases, could be classified (Table 1.1 and 5.3), suggests that epitopes are shared across individuals, which would contradict universal epitope randomness.

Similarly to antibody profiling studies, B-cell epitope approaches differ extensively with respect to the used methodology [177], which explains why B-cell epitope prediction studies suffer from the same lack of comparability as serological diagnostics (Section 9.2.3). In fact, the increase of machine learning methods (Section 1.7.3) for epitope prediction purposes crucially relies on correct B-cell epitope predictions method, since machine learning can only develop new rules [303] based on preexisting knowledge. In this respect, it should be noted that most support vector machines give binary answers[18] [177]. Using signal intensity distributions, as done in this work, might do more justice to the data and would abolish the need for a discrete representation of B-cell epitopes.

[17] A similar finding has been made for protein-peptide docking programs [168].
[18] Multi-class support vector machine approaches also exist [276].

9.4 Assessing the specificity of antibody-peptide reactivity data

While for both unbiased and biased mixtures, the nonlinear interaction of antibody and peptide sequence lays the basis for the simulated binding signals, ensemble properties of highly diverse antibody mixtures cause the genesis of binding signals of unbiased mixtures to be a linear phenomenon.

For both kinds of mixtures, simulated antibody-peptide binding was shown to be non-bijective (Section A.8.1, Equation 4.1): given a binding signal, it is impossible to infer the diversity of the mixture[19]. Ensemble properties add to the non-bijectivity of the Equation for signal intensity generation (Equation 4.1) by abolishing sequence specificity and establishing amino acid composition specificity (Sections 4.3 and A.8.2). As to the experimental setting, the recording of signal intensities, allowed to range continuously between 1 an 65000 [245], is a dimension reduction of high-dimensional binding events in the process of which information is certainly lost.

Even though information is most likely lost in the process of signal intensity recording, numerous studies have shown the potential of antibody binding profiles to discriminate serum samples of different disease stages (Table 1.1). To a certain degree, these profiles even mirror antibody diversity (Section 9.1.2). However, (i) it is incompletely understood how much of the classifiable information of the antibody mixture is actually captured by antibody-peptide binding profiles. (ii) It is also unknown how the classifiability of sera depends on the manufacturer[20] (Section 9.2.3) and the peptide library; removing the highly discriminatory peptides from the data set greatly diminishes classification accuracy (Table S.1) thus calling the diagnostic versatility of random-sequence peptide arrays into question [210]. (iii) Admittedly, signals were found to change upon acute infections (Table 1.1), but to what extent are they disease-specific[21] [210]? (iv) In fact, what do the terms *healthy* and *diseased* signify for antibody binding profiles? Is there a such a thing as a signature of health? Concepts such as *unbiased* mixture and *antibody dominance* as well as the analysis of baseline studies (SHS) are indispensable first steps for answering these questions. Therefore, in order to investigate how and if antibody profile specificity can emerge, the interplay of background serum antibodies [290] and dominant antibodies has to be studied [210].

[19] Concededly, determining the predictive performance of the mixture's signal intensity profile provides one with a rough idea of its diversity (Section 9.1.2).

[20] In the proposed model (Section 4.2), assigned AAWS ($\vec{h}$) are binding priorities, which are unilaterally coded in the amino acids: antibody binding is modeled to be sequence-specific but not amino acid-specific: antibodies cannot be, for example, alanine-affine or averse. A positive component in position two of the antibody sequence impacts the simulated peptide signal intensity positively irrespective of the amino acid in position two of the peptide. Therefore, in order to assess how good an approximation the suggested model is, the statistical properties of antibody binding as well as the immunogenicity of amino acids have to be further investigated. The fact that antigenicity propensity scales for epitope prediction [288] do not perform better than random models [172] (Section 9.3), does not exclude an involvement of amino acid preferences in antibody-peptide binding.

[21] The isolation of the signal of dominant antibodies (Section 4.3.5) may be helpful to study the disease specificity of signals.

9.5 Conclusion

This work shows that the genesis of antibody binding profiles differs with respect to low and high diversity antibody mixtures. Due to ensemble properties of unbiased mixtures, a peptide library's amino acid composition alone is sufficient for binding prediction. In this respect, the agreement of experimental data with model predictions was found to be high as evidenced by both relatively high predictive performance values of mixtures of healthy individuals and a high consistency of AAWS and signal intensities across individuals. The concept of the unbiased mixture is crucial to this work as it offers a deeper understanding of the genesis of antibody binding profiles. Mixtures of healthy individuals were conjectured to comply best with the properties of unbiased mixtures. Furthermore, even in cases of low predictive performance values, the consistency of AAWS was—thanks to their suggested noise resistance—high across individuals, microarray batches, peptide libraries and array printing techniques. Thus, even though the mathematical model used is minimal in its assumptions, it is able to reflect major traits of antibody-peptide reactivity data.

It could be argued that the predictive performance represents under certain assumptions a measure of antibody diversity. However, the direct correspondence of in silico and in vitro antibody diversity with respect to predictive performance is a matter of further investigations.

Serological diagnostics demand both a high degree of consistency across individuals of a given group and a certain degree of difference with respect to any other. The mathematical model predicts that *biased* mixtures dominated by antibodies with similar binding behavior implement this premise.

Furthermore, it became apparent that, in addition to total antibody concentration and antibody dominance, classifiability may depend on the technology used. Therefore, from a biological point of view, measured differences in signal intensity profiles, can, at the current state of technology, be solely interpreted in relative terms (signal intensity, rank and AAWS differences). Absolute signal intensities are of technological nature. Understanding how both technology and antibody mixture together give rise to signal intensities, would open new possibilities for standardizations and quality control as well as the assessment of the specificity of antibody binding profiles [210, 238, 245, 246, 305].

As to B-cell epitope mapping, unbiased mixtures were discovered as an important special case for which amino-acid scale prediction of peptide binding is justified. For other cases, alternative methods have to be sought.

This thesis indicates that a knowledge of both a polyclonal mixture's diversity and composition is essential for the interpretation of antibody binding profiles with respect to both serological diagnostics and B-cell epitope mapping. The statistical binding properties of antibodies merit further study to consolidate these conclusions.

The importance of understanding the emergence of ensemble properties by building quantitative models has been recently indicated by Mora and colleagues [97]. This work presents a new framework for investigating antibody-peptide reactivity data in an unbiased fashion. Thanks to its minimal assumptions approach, the model is a priori applicable to a wide range of questions involving the binding of protein mixtures.

Appendix A

A.1 PLSR: Extended mathematical background

The following extended mathematical background on PLSR is based on articles by Wold, Burnham and colleagues [270, 306].

Consider a dataset where k variables, $\vec{x} = (x_1, x_2, \ldots, x_k)$, are measured as predictor variables and some corresponding response variables $\vec{y} = (y_1, y_2, \ldots, y_k)$. The assumption is that both $\vec{x}$ and $\vec{y}$ have a common underlying latent structure such that the process under observation is actually driven by a set of $a \leq k$ latent variables (also termed components) $\vec{z} = (z_1, z_2, \ldots, z_a)$. These variables are not observable, but their influence can be seen in the measured variables $\vec{x}$. Their relationship is modeled by:

$$\vec{x} = \vec{z}\mathbf{P} + \vec{\varepsilon} \tag{S.1}$$

where $\vec{z}$ is of dimension $1 \times a$, $\mathbf{P}$ of dimension $a \times k$, and $\vec{\varepsilon}$ of dimension $1 \times k$. The last term in the model, $\vec{\varepsilon}$, is considered to be random error. Since $\vec{z}$ is unobservable and $\mathbf{P}$ is unknown, $\vec{z}$ is not identifiable in Equation S.1. In fact, the same values for $\vec{x}$ would arise if $\vec{z}$ and $\mathbf{P}$ are, respectively, replaced with $\vec{z}^* = \vec{z} \times \mathbf{C}$ and $\mathbf{P}^* = \mathbf{C}^{-1} \times \mathbf{P}$, where $\mathbf{C}$ is any non-singular $a \times a$ matrix. Thus, the model is more commonly given as:

$$\vec{x} = \vec{t} \times \mathbf{P} + \vec{\varepsilon} \tag{S.2}$$

where $\vec{t}$ is understood to be some transform $\vec{z}\mathbf{C}$ of the actual latent variables $\vec{z}$. The transformation of $\vec{z}$ to $\vec{t}$ is simply a change of basis so that the points in $\vec{t}$ would lie in the same vector space as those in $\vec{z}$ but expressed in a different basis. In general, the actual latent variables are not as important as the overall space they generate. Therefore, any basis, $\vec{t}$, will be sufficient to define this space. For a given set of n observations following Equation S.2, the model can be written as:

$$\mathbf{X} = \mathbf{TP} + \mathbf{E}, \tag{S.3}$$

where $\mathbf{X}$ is $n \times k$, $\mathbf{T}$ is $n \times a$ and $\mathbf{E}$ is $n \times k$ [306]. In analogy to Equation S.3:

$$\mathbf{Y} = \mathbf{TQ} + \mathbf{F}, \tag{S.4}$$

where again $\mathbf{F}$ is assumed to be random error. $\mathbf{T}$ are also called x-scores. They are predictors of $\mathbf{Y}$ and also model $\mathbf{X}$ (Equations S.3 and S.4).

Appendix B

A.2 Kernel density estimates of monoclonal and serum signal intensity profiles

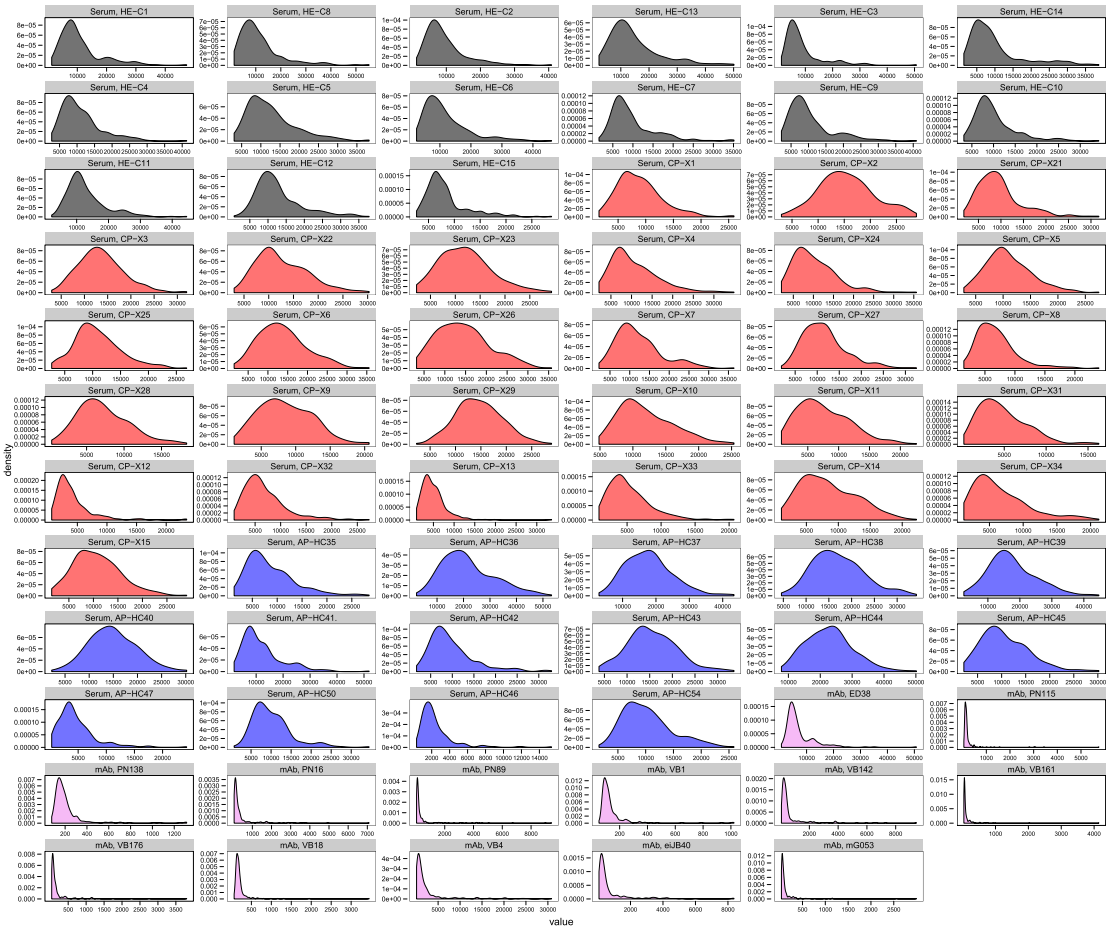


Figure S.1: Gaussian kernel density estimates of signal intensity profiles of human monoclonal IgG (Section 3.5.9) and murine serum IgM antibodies (BALB/c, Mouse study, Section 3.5.8) incubated on the $J^{255}_{\text{14-mer}}$ library are shown. Densities of serum samples of the Mouse study are color-coded by stage of infection (black: *HE*, red: *AP*, blue: *CP*). Panels are named according to the convention: Serum/mAb-(Stage of immune response)-Serum/mAb label.

Appendix C

A.3 Principal component analysis of IgM signal intensity profiles and their ranks

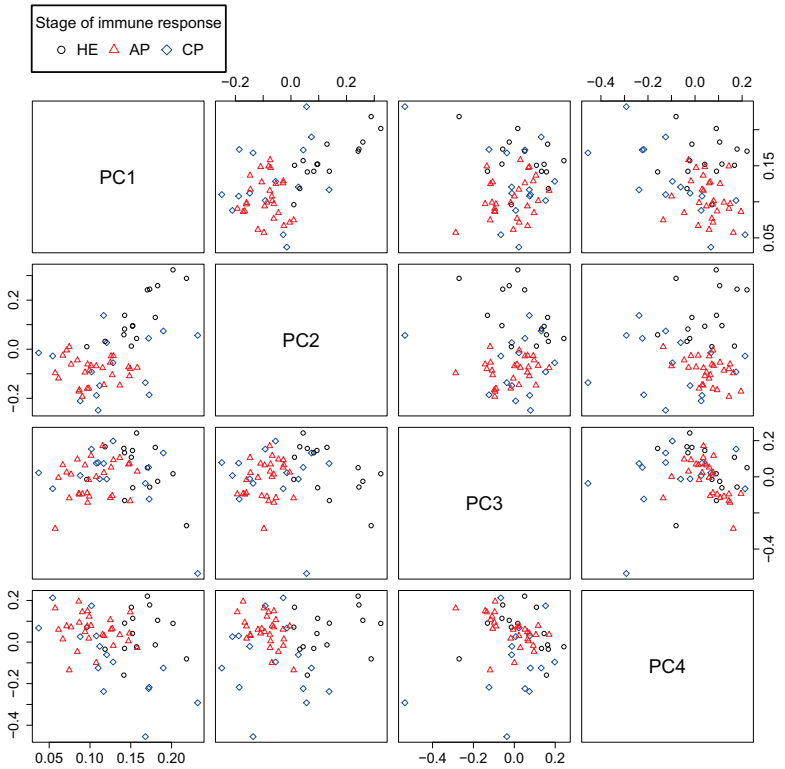


Figure S.2: Stages of immune response differ in their IgM signal intensity profiles. Principal component analysis was applied to the 255×58 signal intensity matrix (255 analyzed peptide signal intensities of library $J^{255}_{\text{14-mer}}$ times 58 BALB/c samples of the Mouse study, Section 3.5.8). The loadings of the first 4 principal components are shown. The first two principal components (PC1, PC2) separate *HE* and diseased (*AP*, *CP*) mice whereas the first and the fourth principal component (PC1, PC4) tend to separate *AP* and *CP* samples. Four components explain (PC1–PC4) 83% of the variance in the data. Sample numbers for the respective groups are: *HE*, 15; *AP*, 28; *CP*, 15.

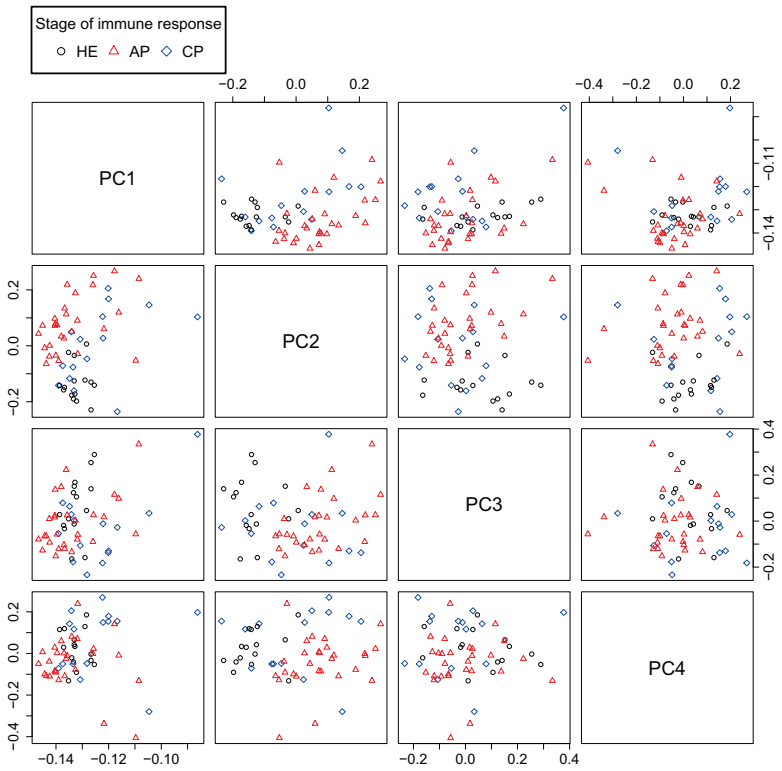


Figure S.3: Stages of immune response differ in their ranks of IgM signal intensity profiles. Principal component analysis was applied to the 255×58 rank matrix (ranks of 255 analyzed peptide signal intensities of library $J^{255}_{\text{14-mer}}$ times 58 BALB/c samples of the Mouse study, Section 3.5.8). The first two principal components (PC1, PC2) separate *HE* and diseased (*AP*, *CP*) mice whereas the second and the fourth principal component (PC2, PC4) tend to separate *AP* and *CP* samples. Four components (PC1–PC4) explain 81% of the variance in the data. Sample numbers for the respective groups are: *HE*, 15; *AP*, 28; *CP*, 15.

A.4 P-SVM classification results after removal of selected peptides

IgM signal intensity profiles					
Subproblem	BACC [%]	Sensitivity [%]	Specificity [%]	Significance (p-value)	Number of removed peptides
HE–AP	89.5	85.7	93.3	0	6
HE–CP	63.4	40.0	86.7	0.072	9
AP–CP	57.0	53.3	60.7	0.108	11
Ranks					
Subproblem	BACC [%]	Sensitivity [%]	Specificity [%]	Significance (p-value)	Number of removed peptides
HE–AP	92.9	85.7	100	0	6
HE–CP	36.7	53.3	20.0	1	15
AP–CP	36.7	73.3	0	1	6

Table S.1: Assessment of the P-SVM balanced classification accuracy (BACC) for both IgM signal intensity profiles and their ranks of subproblems of the Mouse study (BALB/c, Section 3.5.8, Figure 3.1) after removal of previously selected peptides. For each subproblem, prior to nested cross-validation with P-SVM, unique peptides selected for optimal classification results (Table 5.3) were removed from the data set (Section 3.8.2). The BACC is highest for the subproblem *HE–AP* and lowest for *AP–CP*. Signal intensity profiles were determined with the $J^{255}_{14\text{-mer}}$ library. Sample numbers for the respective groups are: *HE*, 15; *AP*, 28; *CP*, 15.

Appendix D

A.5 Secondary-antibody correction of signal intensity profiles

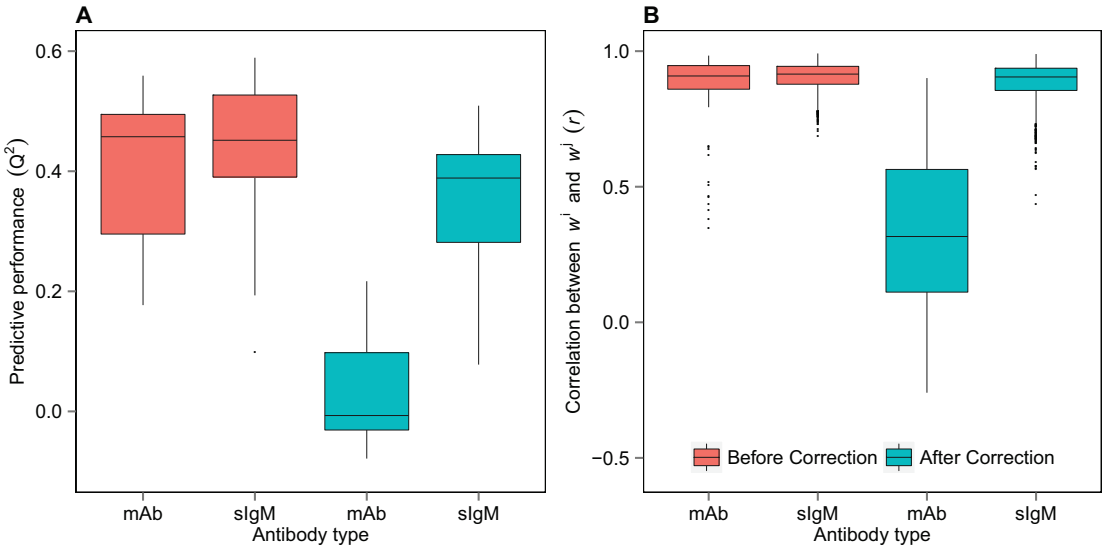


Figure S.4: Assessment of predictive performance values before and after secondary-antibody correction (Mouse study, BALB/c, Section 3.5.8). (A) The predictive performance values (Q^2) were calculated for monoclonal (mAb, Section 3.5.9) and serum IgM (sIgM) antibody binding profiles before (red) and after (blue) correction of the measured log-transformed signal intensities by removal of the polyclonal secondary antibody-correlated signals using PLSR (Section 3.4.2). (B) The pairwise Pearson correlation (r) of the corresponding AAWS $\vec{w}^j$ are shown. For the two statistical measures, signal correction entails a significant decrease in the mAb median, whereas sIgM medians remain largely unchanged. Sample numbers for the respective groups are: mAb, 13; sIgM, 58. Antibody binding profiles were measured with the $J^{255}_{14\text{-mer}}$ library. Corresponding AAWS ($\vec{w}^j$) were determined with Equation 4.8.

Appendix E

A.6 Assessment of the consistency of AAWS across microarray batches, manufacturers and species

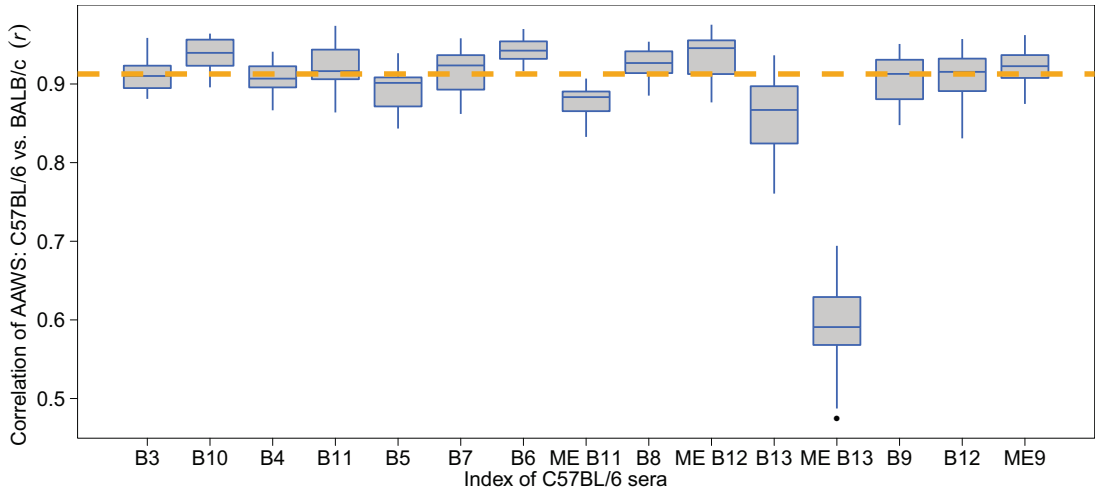


Figure S.5: Mouse study: Comparison between AAWS of BALB/c and C57BL/6 mice shows that AAWS are mostly similar between the two mouse strains. The correlation between AAWS of healthy BALB/c and healthy C57BL/6 mice is shown (overall median Pearson correlation coefficient: $r = 0.91$, orange dashed line). Antibody binding profiles were determined using the $J^{255}_{14\text{-mer}}$ library. Corresponding AAWS were calculated using Equation 4.8. Number of serum samples: BALB/c (healthy), 15; C57BL/6 (healthy), 15 (Section 3.5.8).

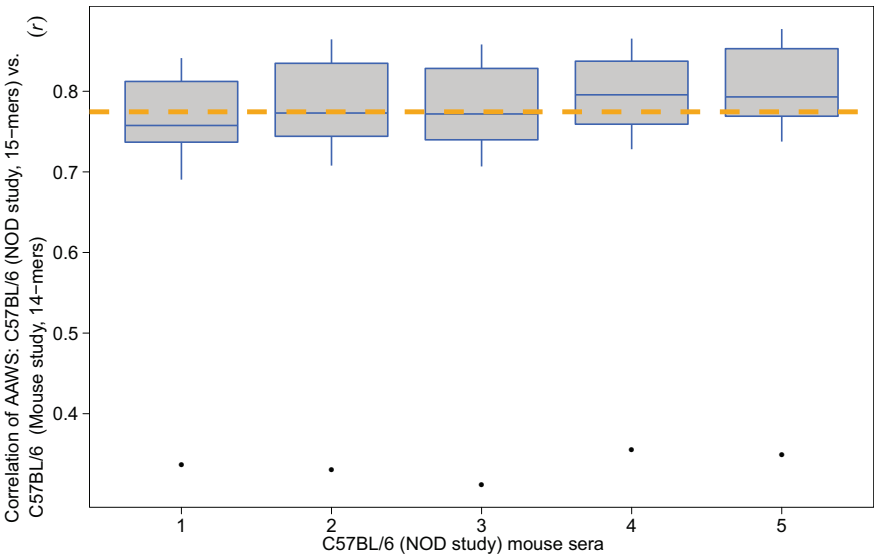


Figure S.6: NOD study, Mouse study: The correlation between AAWS of C57BL/6 mice of NOD and Mouse study is high (overall median correlation coefficients: $r = 0.77$, orange dashed line). (Note that in order to be able to compare AAWS of both studies, the weight of cysteine had to be removed from all Mouse study-AAWS vectors (Table 3.2).) Antibody binding profiles were determined using the libraries $J^{255}_{\text{14-mer}}$ (Mouse study) and $J^{3626}_{\text{15-mer}}$ (NOD study). AAWS were determined using Equation 4.8. Number of serum samples: C57BL/6 (healthy, Mouse study, Section 3.5.8), 15; C57BL/6 (healthy, NOD study, Section 3.5.7), 5.

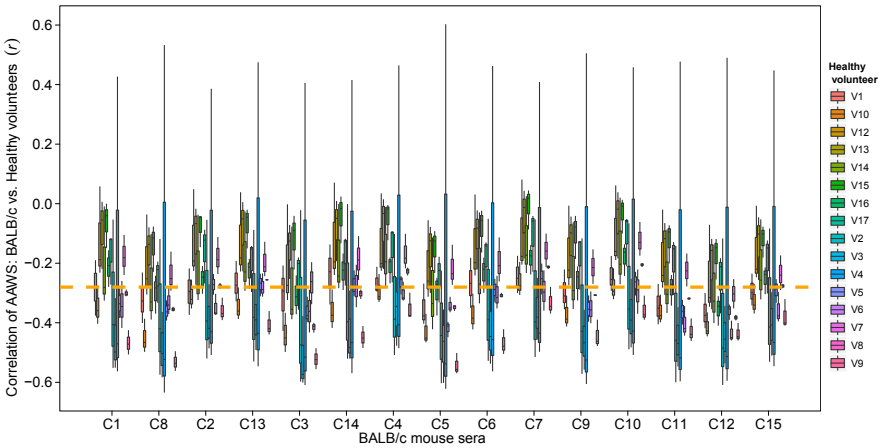


Figure S.7: Mouse study, Slovenian healthy study: the correlation between AAWS of healthy BALB/c mice (MS) and healthy human volunteers (SHS) is poor (overall median correlation coefficient: $r = -0.28$, orange dashed line). (Note that in order to be able to compare AAWS of the MS with AAWS of the SHS, the weight of cysteine had to be removed from all Mouse Study-AAWS vectors (Table 3.2).) Antibody binding profiles were determined using the peptide libraries $J^{255}_{\text{14-mer}}$ (MS) and $J^{3418}_{\text{15-mer}}$ (SHS). AAWS were determined using Equation 4.8. Number of serum samples: BALB/c (healthy, MS, Section 3.5.8), 15; human healthy volunteers (SHS, Section 3.5.1), 48.

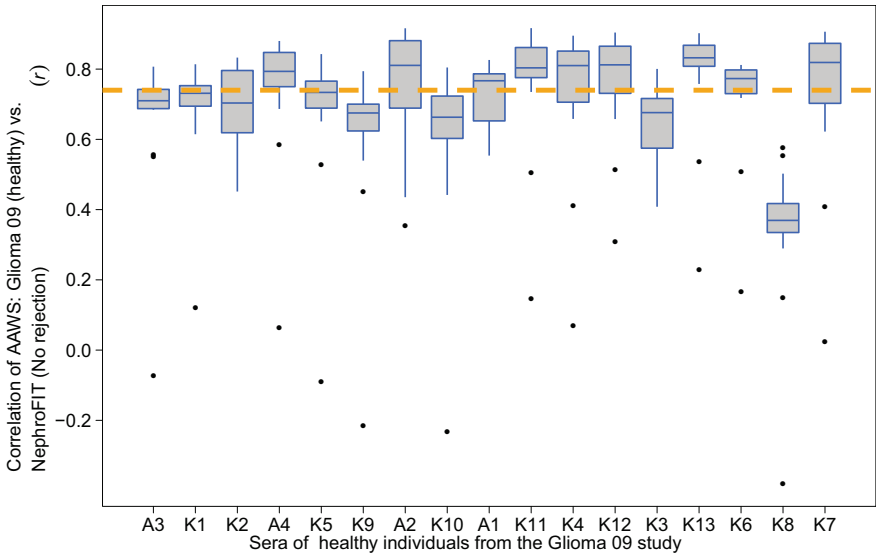


Figure S.8: Glioma 09 study, NephroFIT study: the correlation of AAWS between healthy individuals from the Glioma 09 study and "No rejection" individuals from the NephroFIT study is high (overall median correlation coefficient: $r = 0.74$, orange dashed line). Antibody binding profiles were determined using the libraries $J^{3352}_{\text{15-mer}}$ (Glioma 09 study) and $J^{3418}_{\text{15-mer}}$ (NephroFIT Study). AAWS were determined using Equation 4.8. Number of serum samples: "No rejection" (NephroFIT study, Section 3.5.4), 14; healthy humans (Glioma 09 study, Section 3.5.2), 17.

	Median correlation coefficient of matched pairs (r)	Median correlation coefficient of all possible pairs (r)
AAWS	0.86	0.77
Signal intensity profiles	0.60	0.44

Table S.2: Glioma 08 study, Glioma 09 study: AAWS and signal intensity profiles of matched pairs of both Glioma studies are higher correlated than non-matched ones. AAWS show higher correlations than signal intensity profiles. Serum samples of both studies were matched: 48 samples were assessed in *both* of the studies. Their signal intensities and AAWS were determined and the median of 48 Pearson correlation coefficients calculated (1$^{\text{st}}$ column). As a control, signal intensity profiles and AAWS of matching samples were correlated in a pairwise fashion: sample 1 of Glioma 08 with samples 1 through 48 of Glioma 09, sample 2 of Glioma 08 with samples 1 through 48 of Glioma 09 and so forth such that the median of all pairwise Pearson correlation coefficients could be determined (2$^{\text{nd}}$ column). The correlation of signal intensity profiles of the blanks of both studies was found to be low: $r = 0.22$. As the Glioma 08 ($J^{3418}_{\text{15-mer}}$) and the Glioma 09 studies ($J^{3352}_{\text{15-mer}}$) were done on different peptide libraries (Table 3.1), only signal intensities of matching peptides ($n = 3352$) were used for correlation. Signal intensity profiles were not normalized.

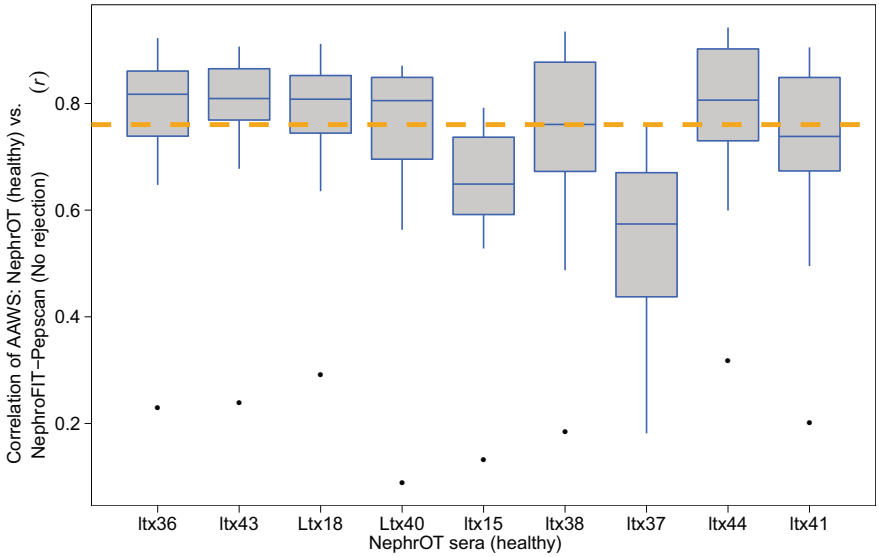


Figure S.9: NephrOT, NephroFIT-Pepscan study: correlation between AAWS of healthy individuals from the NephrOT study and "No rejection" individuals from the NephroFIT-Pepscan study is high (overall median correlation coefficient: $r = 0.76$, orange dashed line). The overall median coefficient for raw (unnormalized, not log-transformed) signal intensities is $r = 0.23$. The correlation of signal intensities of respective blanks is $r = 0.35$. Antibody binding profiles were determined with the library $P^{942}_{15\text{-mer}}$. AAWS were determined using Equation 4.8. Number of serum samples: "No rejection" (NephroFIT-Pepscan study, Section 3.5.5), 13; healthy humans (NephrOT study, Section 3.5.6), 9.

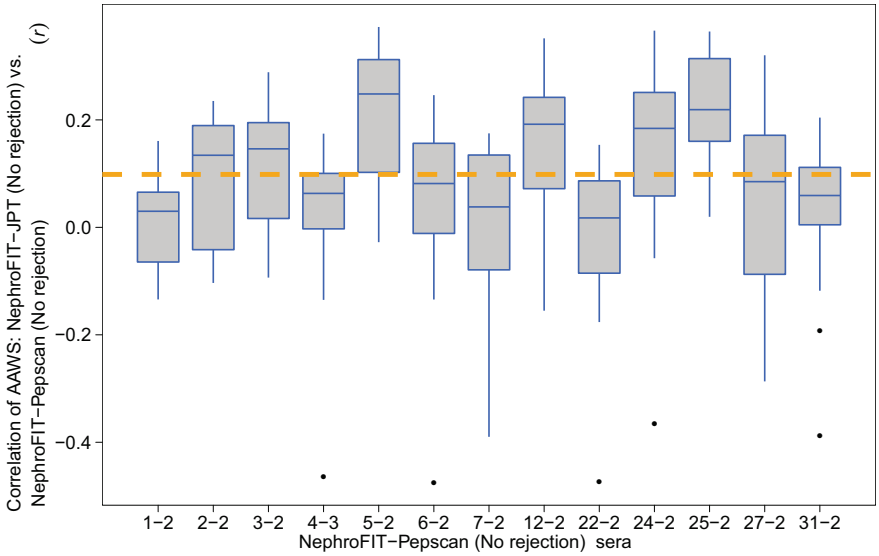


Figure S.10: NephroFIT, NephroFIT-Pepscan study: correlation between AAWS of the NephroFIT-Pepscan and NephroFIT studies (both "No rejection") is low (overall median correlation coefficient: $r = 0.10$, orange dashed line). Antibody binding profiles were determined using the libraries $P^{942}_{15\text{-mer}}$ (NephroFIT-Pepscan) and $J^{3418}_{15\text{-mer}}$ (NephroFIT Study). AAWS were determined using Equation 4.8. Number of serum samples: "No rejection" (NephroFIT-Pepscan study, Section 3.5.5), 13; "No rejection" (NephroFIT study, Section 3.5.4), 14.

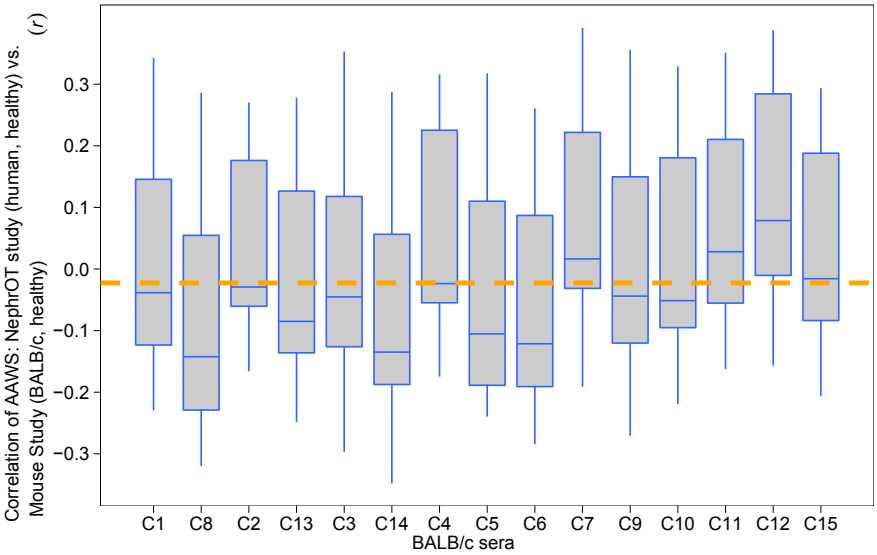


Figure S.11: Mouse study (BALB/c), NephrOT study: correlation between AAWS of healthy BALB/c mice (Mouse study) and healthy individuals from the NephrOT study is low (overall median correlation coefficient: $r = -0.02$, orange dashed line). Antibody binding profiles were determined using the libraries $J^{255}_{14\text{-mer}}$ (Mouse study) and $P^{942}_{15\text{-mer}}$ (NephrOT study). AAWS were determined using Equation 4.8. Number of serum samples: BALB/c (healthy, Mouse study, Section 3.5.8), 15; healthy humans (NephrOT study, Section 3.5.6), 9.

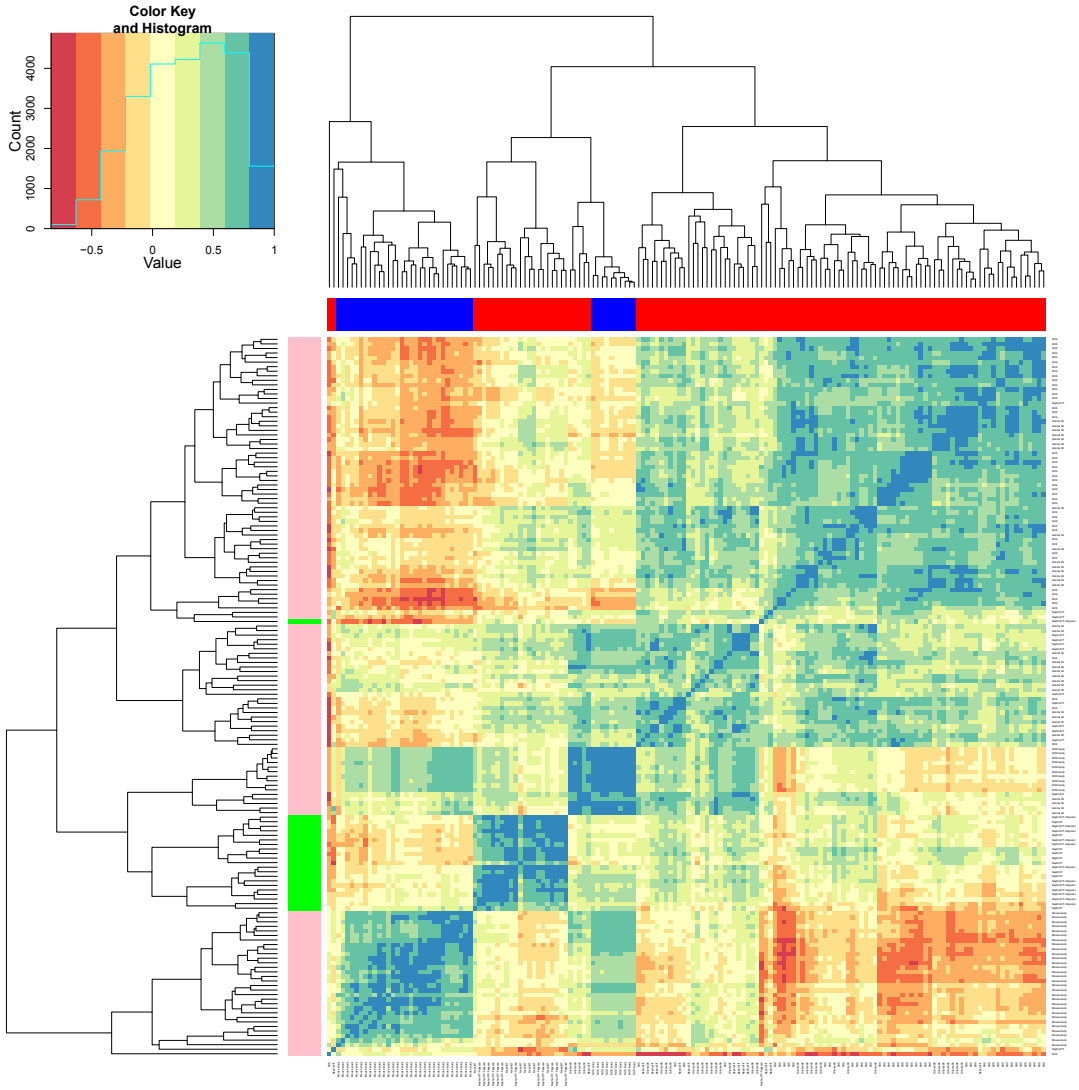


Figure S.12: Heatmap of AAWS of healthy individuals of all experimental studies assessed in this thesis shows Spearman-based clustering of AAWS by species (mouse: blue, human: red) and manufacturer (Pepscan: green, JPT: pink). Number of total AAWS clustered: 158. (For both the NephroFIT and the NephroFIT-Pepscan study, "No rejection" individuals were assessed in this Figure, because no more than two samples of healthy controls were incubated in these studies). AAWS of 13-mer libraries were excluded from this thesis. The weight of cysteine (C) was removed from all AAWS of cysteine-containing libraries ($J^{255}_{\text{14-mer}}$, Table 3.1). Antibody binding profiles were measured with the respective libraries and AAWS were determined with Equation 4.8. The heatmap was built by Spearman-based correlation as detailed in Section 3.9.2.

Appendix F

A.7 PCA and P-SVM nested cross-validation of antibody binding profiles of unbiased mixtures differing by total antibody concentration

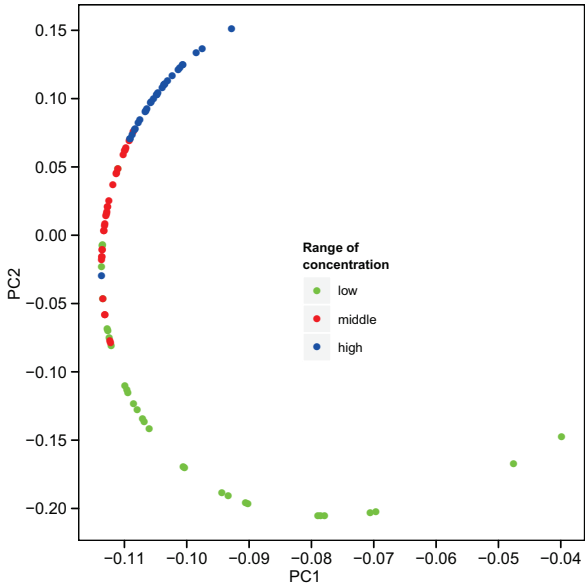


Figure S.13: Antibody binding profiles simulated with different total antibody concentrations cluster when applied to PCA. Signal intensity profiles simulated for Figure 6.9 are shown in the space spanned by first two principal components (PC1, PC2). Together, PC1 and PC2 explain nearly 100% of the variance in the data. Please refer to Figure 6.9 for simulation details.

Simulated signal intensity profiles				
Subproblem	BACC [%]	Sensitivity [%]	Specificity [%]	Significance (p-value)
Low–Middle	83.3	93.3	73.3	0
Low–High	91.7	96.7	86.7	0
Middle–High	88.7	96.7	80.7	0

Table S.3: Assessment of the P-SVM balanced classification accuracy (BACC) for signal intensity profiles simulated with different antibody concentrations. Signal intensity profiles are those simulated for Figure 6.9. Please refer to Figure 6.9 for simulation details.

Appendix G

A.8 A minimal model of antibody-peptide binding: further mathematical analyses

A.8.1 Signal intensity simulation is a non-bijective process

Let f be a function of the vectors $\vec{a}^k$ and $\vec{p}^i$, and of the scalar n_{Ab}.
$f : \mathbb{R}^l \times \mathbb{R}^l \times \mathbb{N}_{\geq 1} \to (0, 1)$, with

$$f(\vec{a}^k, \vec{p}^i, n_{\mathrm{Ab}}) = S_i = \frac{\sum_{k=1}^{n_{\mathrm{Ab}}} [\mathrm{Ab}]_k K_{i,k}}{1 + \sum_{k=1}^{n_{\mathrm{Ab}}} [\mathrm{Ab}]_k K_{i,k}}. \tag{S.5}$$

Let A_1 and A_2 be two antibody repertoires of cardinality 2, with $\vec{a}^1, \vec{a}^2 \in \mathrm{A}_1$ and $\vec{a}^3, \vec{a}^4 \in \mathrm{A}_2$, and $\mathrm{AM}_1 = \mathrm{A}_1, \mathrm{AM}_2 = \mathrm{A}_2$. Also, let $\vec{a}^3 = \vec{a}^1 + \vec{a}^2, \vec{a}^4 = \vec{0}$ and $\vec{p}^i = (0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1)^{\mathrm{T}}$. Then, it follows:

$$f(\mathrm{AM}_1(\vec{a}^1, \vec{a}^2, n_{\mathrm{Ab}_1}), \vec{p}^i) = S_i = f(\mathrm{AM}_2(\vec{a}^3, \vec{a}^4, n_{\mathrm{Ab}_2}), \vec{p}^i). \tag{S.6}$$

Therefore, f is not injective[22] and thereby not bijective. The simulation of signal intensities is a process which does not preserve the entire information (e.g. antibody repertoire size) of the input space.

A.8.2 Unbiased mixtures reduce the dimensionality of the signal intensity space

Signal intensity profiles of unbiased mixtures are predictable by AAWS (Section 4). This finding implies that signal intensity profiles of unbiased mixtures span a dimensionally reduced signal intensity space since peptides with different antibody sequences but the same amino acid composition are predicted to yield the same signal intensity.

Given an unbiased mixture, the number of unique peptide sequences—with respect to amino acid order—drawn from an arbitrary amino acid alphabet AA is reduced $-\log \Gamma_n^k / n^k$-fold.

$$\Gamma_n^k / n^k = \frac{1}{n^k} \binom{n + k - 1}{k}, \tag{S.7}$$

[22] If n_{Ab} were 1 and the signal intensity profile with the respective peptide sequences and assigned AAWS ($\vec{h}$) were known, then the antibody sequence of the one antibody could be largely recovered thanks to the approximation obtained by the Taylor series expansion of the exponential function, $\exp(x) = 1 + x$, and linear regression.

where $n = \#\mathrm{AA}$ is the size of the amino acid alphabet and l is the peptide sequence length.

Let l amino acids be drawn from a set of n different amino acids with repetitions allowed and the order of drawing is relevant. The number of possible amino acid sequences (peptides) is then n^k. Therefore, the aim is to show that the numerator of Equation S.7 yields the number of all possible combinations of peptide sequences (i) if, for one sequence, k out of n amino acids are drawn randomly, (ii) if the order of the drawn amino acids is of no importance and (iii) drawing repetitively the same amino acid is allowed. If drawing *repetitively* the same amino acid was not allowed, the above assumptions would lead to: $\binom{n}{k}$. Using the ansatz to mapping bijectively the "repetition" case onto the "non-repetition" case yields Equation S.7 [307].

Setting the size of the amino acid alphabet to $\#\mathrm{AA} = 20$ and the peptide sequence length l to 14 (Figure 4.2), Equation S.7 yields 4.99×10^{-10}, which signifies that the number of unique sequences with respect to amino acid order is reduced $\approx 10^{10}$-fold.

A.8.3 Derivation of the isolation of the signal of dominant antibodies from a biased mixture's signal

The approximative isolation of the signal of the dominant antibodies $S_{i,\mathrm{D}}$ (Equation 4.9) can be derived as follows.

$$S_{i,\text{U-D}} = \frac{\sum_{k=1}^{n_{\mathrm{Ab,U\text{-}D}}} [\mathrm{Ab}]_k K_{i,k}}{1 + \sum_{k=1}^{n_{\mathrm{Ab,U\text{-}D}}} [\mathrm{Ab}]_k K_{i,k}} \tag{S.8}$$

$$S_{i,\text{U-D}} = \frac{\sum_{k=1}^{n_{\mathrm{Ab,U\text{-}D}} - n_{\mathrm{Ab,D}}} [\mathrm{Ab}]_k K_{i,k} + \sum_{k=1}^{n_{\mathrm{Ab,D}}} [\mathrm{Ab}]_k K_{i,k}}{1 + \sum_{k=1}^{n_{\mathrm{Ab,U\text{-}D}} - n_{\mathrm{Ab,D}}} [\mathrm{Ab}]_k K_{i,k} + \sum_{k=1}^{n_{\mathrm{Ab,D}}} [\mathrm{Ab}]_k K_{i,k}} \tag{S.9}$$

$$S_{i,\text{U-D}} \approx \frac{s_{i,\mathrm{U}} + s_{i,\mathrm{D}}}{1 + s_{i,\mathrm{U}} + s_{i,\mathrm{D}}} \tag{S.10}$$

$$(1 + s_{i,\mathrm{U}} + s_{i,\mathrm{D}}) S_{i,\text{U-D}} = s_{i,\mathrm{U}} + s_{i,\mathrm{D}} \tag{S.11}$$

$$(1 + s_{i,\mathrm{U}}) S_{i,\text{U-D}} + s_{i,\mathrm{D}} S_{i,\text{U-D}} = s_{i,\mathrm{U}} + s_{i,\mathrm{D}} \tag{S.12}$$

$$s_{i,\mathrm{D}} (S_{i,\text{U-D}} - 1) = s_{i,\mathrm{U}} - (1 + s_{i,\mathrm{U}}) S_{i,\text{U-D}} \tag{S.13}$$

$$s_{i,\mathrm{D}} = \frac{s_{i,\mathrm{U}} - (1 + s_{i,\mathrm{U}}) S_{i,\text{U-D}}}{S_{i,\text{U-D}} - 1} \tag{S.14}$$

$$s_{i,\mathrm{D}} = \frac{\frac{-S_{i,\mathrm{U}}}{S_{i,\mathrm{U}} - 1} - (1 + \frac{-S_{i,\mathrm{U}}}{S_{i,\mathrm{U}} - 1}) S_{i,\text{U-D}}}{S_{i,\text{U-D}} - 1} \tag{S.15}$$

$$s_{i,\mathrm{D}} = \frac{S_{i,\text{U-D}} \frac{1}{S_{i,\mathrm{U}} - 1} - \frac{S_{i,\mathrm{U}}}{S_{i,\mathrm{U}} - 1}}{S_{i,\text{U-D}} - 1} \tag{S.16}$$

$$s_{i,\mathrm{D}} = \frac{S_{i,\text{U-D}} - S_{i,\mathrm{U}}}{(S_{i,\text{U-D}} - 1)(S_{i,\mathrm{U}} - 1)} \tag{S.17}$$

Bibliography

[1] Javier Mestas and Christopher C. W Hughes. Of Mice and Not Men: Differences Between Mouse and Human Immunology. *The Journal of Immunology*, 172(5):2731–2738, 2004.

[2] C. Janeway, M. J. Shlomchik, and Walport. *Immunobiology*. Garland Science, 6 edition, 2004.

[3] A R Abbas, D Baldwin, Y Ma, W Ouyang, A Gurney, F Martin, S Fong, M van Lookeren Campagne, P Godowski, P M Williams, A C Chan, and H F Clark. Immune response in silico (IRIS): immune-specific genes identified from a compendium of microarray expression data. *Genes and Immunity*, 6(4):319–331, 2005.

[4] Thomas Heams, Philippe Huneman, Guillaume Lecointre, and Marc Silberstein. *Les mondes darwiniens. L'évolution de l'évolution*. Editions Syllepse, 2009.

[5] IR Cohen, U Hershberg, and S Solomon. Antigen-receptor degeneracy and immunological paradigms. *Molecular Immunology*, 40(14-15):996, 993, 2004.

[6] Danielle L Drayton, Shan Liao, Rawad H Mounzer, and Nancy H Ruddle. Lymphoid organ development: from ontogeny to neogenesis. *Nature Immunology*, 7(4):344–353, 2006.

[7] Björn Hartmann, René Riedel, Katharina Jörss, Christoph Loddenkemper, Andreas Steinmeyer, Ulrich Zügel, Magda Babina, Andreas Radbruch, and Margitta Worm. Vitamin D receptor activation improves allergen-triggered eczema in mice. *The Journal of Investigative Dermatology*, 132(2):330–336, 2012.

[8] Ruslan Medzhitov and Charles A Janeway Jr. Innate immune recognition and control of adaptive immune responses. *Seminars in Immunology*, 10(5):351–353, 1998.

[9] William A McEwan, Donna L Mallery, David A Rhodes, John Trowsdale, and Leo C James. Intracellular antibody-mediated immunity and the role of TRIM21. *BioEssays*, 33(11):803–809, 2011.

[10] Thomas Pradeu. *L'immunologie et la définition de l'identité biologique*. Philosophy, l'Université de Paris 1 Panthéon-Sorbonne, 2007.

[11] J A Berzofsky. Intrinsic and extrinsic factors in protein antigenic structure. *Science (New York, N.Y.)*, 229(4717):932–940, 1985.

[12] Marc H. V. van Regenmortel. The Recognition of Proteins and Peptides by Antibodies. *Journal of Immunoassay*, 21(2-3):85–108, 2000.

[13] Sir; Mavis Freeman; A V Jackson; Dora Lush F M Burnet. *The Production of Antibodies. A Review and a Theoretical Discussion.* Melbourne, London, Macmillan and Co. Ltd., 1941.

[14] Sir Frank Macfarlane Burnet. *Self and not-self: Cellular Immunology, 1.* Melbourne University Press, 1969.

[15] Thomas Pradeu and Edgardo D. Carosella. On the definition of a criterion of immunogenicity. *Proceedings of the National Academy of Sciences*, 103(47):17858–17861, 2006.

[16] Rodney E. Langman. The specificity of immunological reactions. *Molecular Immunology*, 37(10): 555–561, 2000.

[17] Gary W. Litman, Jonathan P. Rast, and Sebastian D. Fugmann. The origins of vertebrate adaptive immunity. *Nat Rev Immunol*, 10(8):543–553, 2010.

[18] Zvi Grossman and William E. Paul. Self-tolerance: context dependent tuning of T cell antigen recognition. *Seminars in Immunology*, 12(3):197–203, 2000.

[19] Polly Matzinger. The Danger Model: A Renewed Sense of Self. *Science*, 296(5566):301–305, 2002.

[20] I R Cohen. The cognitive principle challenges clonal selection. *Immunology Today*, 13(11):441–4, 1992.

[21] Arthur M. Silverstein and Noel R. Rose. On the mystique of the immunological self. *Immunological Reviews*, 159(1):197–206, 1997.

[22] Alfred I. Tauber. *The Immune Self: Theory or Metaphor?* Cambridge University Press, reprint edition, 1996.

[23] P Matzinger. Tolerance, danger, and the extended family. *Annual Review of Immunology*, 12: 991–1045, 1994.

[24] NS Greenspan, DA Dacek, and LJ Cooper. Cooperative binding of two antibodies to independent antigens by an Fc-dependent mechanism. *The FASEB Journal*, 3(10):2203–2207, 1989.

[25] M M Morelock, R. Rothlein, S M Bright, M K Robinson, E T Graham, J P Sabo, R. Owens, D J King, S H Norris, and D S Scher. Isotype Choice for Chimeric Antibodies Affects Binding Properties. *Journal of Biological Chemistry*, 269(17):13048–13055, 1994.

[26] Tarun K Dam, Marcela Torres, C. Fred Brewer, and Arturo Casadevall. Isothermal titration calorimetry reveals differential binding thermodynamics of variable region-identical antibodies differing in constant region for a univalent ligand. *Journal of Biological Chemistry*, 283(46): 31366–31370, 2008.

[27] Nicole Wittenbrink. *New perspectives in the evolution of B lymphocytes in germinal centers*. PhD thesis, Humboldt Universität zu Berlin, 2007.

[28] Heinz Penzlin. *Lehrbuch der Tierphysiologie*. Spektrum Akademischer Verlag, 2008.

[29] Bjoern Peters, John Sidney, Phil Bourne, Huynh-Hoa Bui, Soeren Buus, Grace Doh, Ward Fleri, Mitch Kronenberg, Ralph Kubo, Ole Lund, David Nemazee, Julia V Ponomarenko, Muthu Sathiamurthy, Stephen Schoenberger, Scott Stewart, Pamela Surko, Scott Way, Steve Wilson, and Alessandro Sette. The Immune Epitope Database and Analysis Resource: From Vision to Blueprint. *PLoS Biology*, 3(3), 2005.

[30] Julia V Ponomarenko and Philip E Bourne. Antibody-protein interactions: benchmark datasets and prediction tools evaluation. *BMC Structural Biology*, 7(1):64, 2007.

[31] N. S Greenspan. Epitopes, paratopes and other topes : do immunologists know what they are talking about? *Bulletin de l'Institut Pasteur*, 90(4):267–279, 1992.

[32] M H Van Regenmortel. Structural and functional approaches to the study of protein antigenicity. *Immunology Today*, 10(8):266–272, 1989.

[33] D R Davies and G H Cohen. Interactions of protein antigens with antibodies. *Proceedings of the National Academy of Sciences*, 93(1):7–12, 1996.

[34] D J Barlow, M S Edwards, and J M Thornton. Continuous and discontinuous protein antigenic determinants. *Nature*, 322(6081):747–748, 1986.

[35] W G Laver, G M Air, R G Webster, and S J Smith-Gill. Epitopes on protein antigens: misconceptions and realities. *Cell*, 61(4):553–556, 1990.

[36] J A Schroer, T Bender, R J Feldmann, and K J Kim. Mapping epitopes on the insulin molecule using monoclonal antibodies. *European Journal of Immunology*, 13(9):693–700, 1983.

[37] Eric J Sundberg, Roy A Mariuzza, Joel Janin, and Shoshana J. Wodak. Molecular recognition in antibody-antigen complexes. In *Protein Modules and Protein-Protein Interaction*, volume Volume 61, pages 119–160. Academic Press, 2002.

[38] Vered Kunik, Bjoern Peters, and Yanay Ofran. Structural Consensus among Antibodies Defines the Antigen Binding Site. *PLoS Computational Biology*, 8(2), 2012.

[39] T B Lavoie, W N Drohan, and S J Smith-Gill. Experimental analysis by site-directed mutagenesis of somatic mutation effects on affinity and fine specificity in antibodies specific for lysozyme. *Journal of Immunology (Baltimore, Md.: 1950)*, 148(2):503–513, 1992.

[40] Ronald Frank. The SPOT-synthesis technique: Synthetic peptide arrays on membrane supports–principles and applications. *Journal of Immunological Methods*, 267(1):13–26, 2002.

[41] I.Saira Mian, Arthur R. Bradwell, and Arthur J. Olson. Structure, function and properties of antibody binding sites. *Journal of Molecular Biology*, 217(1):133–151, 1991.

[42] S. Sheriff, E W Silverton, E A Padlan, G H Cohen, S J Smith-Gill, B C Finzel, and D R Davies. Three-dimensional structure of an antibody-antigen complex. *Proceedings of the National Academy of Sciences*, 84(22):8075–8079, 1987.

[43] Wesley E. Stites. Protein-Protein Interactions: Interface Structure, Binding Thermodynamics, and Mutational Analysis. *Chem. Rev.*, 97(5):1233–1250, 1997.

[44] E A Kabat, T T Wu, and H Bilofsky. Unusual distributions of amino acids in complementarity-determining (hypervariable) segments of heavy and light chains of immunoglobulins and their possible roles in specificity of antibody-combining sites. *The Journal of Biological Chemistry*, 252 (19):6609–6616, 1977.

[45] E.A. Padlan. On the nature of antibody combining sites: unusual structural features that may confer on these sites an enhanced capacity for binding ligands. *Proteins: Structure, Function, and Bioinformatics*, 7(2):112–124, 1990.

[46] T N Bhat, G A Bentley, T O Fischmann, G Boulot, and R J Poljak. Small rearrangements in structures of Fv and Fab fragments of antibody D1.3 on antigen binding. *Nature*, 347(6292): 483–485, 1990.

[47] D R Davies, E A Padlan, and S Sheriff. Antibody-antigen complexes. *Annual Review of Biochemistry*, 59:439–473, 1990.

[48] Shoshana J Wodak, Joël Janin, Joel Janin, and Shoshana J. Wodak. Structural basis of macromolecular recognition. In *Protein Modules and Protein-Protein Interaction*, volume Volume 61, pages 9–73. Academic Press, 2002.

[49] Q X Hua, S E Shoelson, M Kochoyan, and M A Weiss. Receptor binding redefined by a structural switch in a mutant human insulin. *Nature*, 354(6350):238–241, 1991.

[50] Neil S Greenspan. Cohen's Conjecture, Howard's Hypothesis, and Ptashne's Ptruth: an exploration of the relationship between affinity and specificity. *Trends in Immunology*, 31(4):138–143, 2010.

[51] Bertrand Friguet, Alain F. Chaffotte, Lisa Djavadi-Ohaniance, and Michel E. Goldberg. Measurements of the true affinity constant in solution of antigen-antibody complexes by enzyme-linked immunosorbent assay. *Journal of Immunological Methods*, 77(2):305–319, 1985.

[52] Ibrahim Abdulhalim, Mohammad Zourob, and Akhlesh Lakhtakia. Surface Plasmon Resonance for Biosensing: A Mini-Review. *Electromagnetics*, 28(3):214–242, 2008.

[53] J M Pitarke, V M Silkin, E V Chulkov, and P M Echenique. Theory of surface plasmons and surface-plasmon polaritons. *Reports on Progress in Physics*, 70(1):1–87, 2007.

[54] A. N. Glazer. On the Prevalence of „Nonspecific" Binding at the Specific Binding Sites of Globular Proteins*. *Proceedings of the National Academy of Sciences of the United States of America*, 65(4): 1057–1063, 1970.

[55] Herman N. Eisen and Gregory W. Siskind. Variations in Affinities of Antibodies during the Immune Responses. *Biochemistry*, 3(7):996–1008, 1964.

[56] Gregory A Michaud, Michael Salcius, Fang Zhou, Rhonda Bangham, Jaclyn Bonin, Hong Guo, Michael Snyder, Paul F Predki, and Barry I Schweitzer. Analyzing antibody specificity with whole proteome microarrays. *Nature Biotechnology*, 21(12):1509–1512, 2003.

[57] Paul F Predki, Dawn Mattoon, Rhonda Bangham, Barry Schweitzer, and Gregory Michaud. Protein microarrays: a new tool for profiling antibody cross-reactivity. *Human Antibodies*, 14(1-2):7–15, 2005.

[58] Ulrich Reineke, Claudia Ivascu, Marén Schlief, Christiane Landgraf, Seike Gericke, Grit Zahn, Hanspeter Herzel, Rudolf Volkmer-Engert, and Jens Schneider-Mergener. Identification of distinct antibody epitopes and mimotopes from a peptide array of 5520 randomly generated sequences. *Journal of Immunological Methods*, 267(1):37–51, 2002.

[59] F.F. Richards and WH Konigsberg. Speculations how specific are antibodies? *Immunochemistry*, 10(8):545–553, 1973.

[60] F F Richards, W H Konigsberg, R W Rosenstein, and J M Varga. On the specificity of antibodies. *Science (New York, N.Y.)*, 187(4172):130–137, 1975.

[61] David Schubert, Ann Roman, and Melvin Cohn. Anti-Nucleic Acid Specificities of Mouse Myeloma Immunoglobulins. *, Published online: 10 January 1970; | doi:10.1038/225154a0*, 225(5228):154–158, 1970.

[62] Karl Landsteiner. *The Specificity of Serological Reactions Revised Edition*. Harvard University Press, 1947.

[63] Kai W. Wucherpfennig, Paul M. Allen, Franco Celada, Irun R. Cohen, Rob De Boer, K. Christopher Garcia, Byron Goldstein, Ralph Greenspan, David Hafler, Philip Hodgkin, Erik S. Huseby, David C. Krakauer, David Nemazee, Alan S. Perelson, Clemencia Pinilla, Rol, K. Strong, and Eli E. Sercarz. Polyspecificity of T cell and B cell receptor recognition. *Seminars in Immunology*, 19(4):216–224, 2007.

[64] AR Williamson. Extent and control of antibody diversity. *Biochemical Journal*, 130(2):325, 1972.

[65] Frederik W Wiegel and Alan S Perelson. Some Scaling Principles for the Immune System. *Immunology and Cell Biology*, 82(2):127–131, 2004.

[66] Alan S. Perelson and George F. Oster. Theoretical studies of clonal selection: Minimal antibody repertoire size and reliability of self-non-self discrimination. *Journal of Theoretical Biology*, 81(4): 645–670, 1979.

[67] J. Haimovich and L. Du Pasquier. Specificity of antibodies in amphibian larvae possessing a small number of lymphocytes. *Proceedings of the National Academy of Sciences*, 70(6):1898, 1973.

[68] Alan S. Perelson and Gérard Weisbuch. Immunology for physicists. *Reviews of Modern Physics*, 69 (4):1219, 1997.

[69] R J de Boer and A S Perelson. Size and connectivity as emergent properties of a developing immune network. *Journal of Theoretical Biology*, 149(3):381–424, 1991.

[70] R. M Zinkernagel. Uncertainties- discrepancies in immunology. *Immunological reviews*, 185(1): 103–125, 2002.

[71] N S Greenspan. Affinity, complementarity, cooperativity, and specificity in antibody recognition. *Current Topics in Microbiology and Immunology*, 260:65–85, 2001.

[72] Ilona Mandrika, Peteris Prusis, Sviatlana Yahorava, Kaspars Tars, and Jarl E.S. Wikberg. QSAR of multiple mutated antibodies. *Journal of Molecular Recognition*, 20(2):97–102, 2007.

[73] Ilona Mandrika, Peteris Prusis, Sviatlana Yahorava, Medya Shikhagaie, and Jarl E.S. Wikberg. Proteochemometric modelling of antibody-antigen interactions using SPOT synthesised peptide arrays. *Protein Engineering, Design and Selection*, page gzm022, 2007.

[74] Hedda Wardemann, Sergey Yurasov, Anne Schaefer, James W Young, Eric Meffre, and Michel C Nussenzweig. Predominant autoantibody production by early human B cell precursors. *Science (New York, N.Y.)*, 301(5638):1374–1377, 2003.

[75] Hugo Mouquet, Johannes F. Scheid, Markus J. Zoller, Michelle Krogsgaard, Rene G. Ott, Shetha Shukair, Maxim N. Artyomov, John Pietzsch, Mark Connors, Florencia Pereyra, Bruce D. Walker, David D. Ho, Patrick C. Wilson, Michael S. Seaman, Herman N. Eisen, Arup K. Chakraborty, Thomas J. Hope, Jeffrey V. Ravetch, Hedda Wardemann, and Michel C. Nussenzweig. Polyreactivity increases the apparent affinity of anti-HIV antibodies by heteroligation. *Nature*, 467(7315):591–595, 2010.

[76] Thomas Tiller, Makoto Tsuiji, Sergey Yurasov, Klara Velinzon, Michel C. Nussenzweig, and Hedda Wardemann. Autoreactivity in human IgG+ memory B cells. *Immunity*, 26(2):205–213, 2007.

[77] David W. Talmage. Immunological Specificity: Unique combinations of selected natural globulins provide an alternative to the classical concept. *Science*, 129(3364):1643–1648, 1959.

[78] Arturo Casadevall and Liise-anne Pirofski. A new synthesis for antibody-mediated immunity. *Nat Immunol*, 13(1):21–28, 2012.

[79] C. Berek, G. M. Griffiths, and C. Milstein. Molecular events during maturation of the immune response to oxazolone. *Nature*, 316(6027):412–418, 1985.

[80] Jean-Claude Weill and Claude-Agnès Reynaud. Rearrangement/hypermutation/gene conversion: when, where and why? *Immunology Today*, 17(2):92–97, 1996.

[81] Claudia Berek and César Milstein. The dynamic nature of the antibody repertoire. *Immunological Reviews*, 105:5–26, 1988.

[82] Michael S Neuberger and César Milstein. Somatic hypermutation. *Current Opinion in Immunology*, 7(2):248–254, 1995.

[83] Susumu Tonegawa. Somatic generation of antibody diversity. *Nature*, 302(5909):575–581, 1983.

[84] J.C. Almagro, I. Hernández, M.C. Ramírez, and E. Vargas-Madrazo. Structural differences between the repertoires of mouse and human germline genes and their evolutionary implications. *Immunogenetics*, 47(5):355–363, 1998.

[85] I C M MacLennan. Germinal Centers. *Annual Review of Immunology*, 12(1):117–139, 1994.

[86] D. McKean, K. Huppi, M. Bell, L. Staudt, W. Gerhard, and M. Weigert. Generation of antibody diversity in the immune response of BALB/c mice to influenza virus hemagglutinin. *Proceedings of the National Academy of Sciences*, 81(10):3180–3184, 1984.

[87] NS Levy, UV Malipiero, SG Lebecque, and PJ Gearhart. Early onset of somatic mutation in immunoglobulin VH genes during the primary immune response. *J. Exp. Med.*, 169(6):2007–2019, 1989.

[88] Nancy S. Longo and Peter E. Lipsky. Why do B cells mutate their immunoglobulin receptors? *Trends in Immunology*, 27(8):374–380, 2006.

[89] A G Betz, C Rada, R Pannell, C Milstein, and M S Neuberger. Passenger transgenes reveal intrinsic specificity of the antibody hypermutation mechanism: clustering, polarity, and specific hot spots. *Proceedings of the National Academy of Sciences of the United States of America*, 90(6):2385–2388, 1993.

[90] Armin A Weiser, Nicole Wittenbrink, Lei Zhang, Andrej I Schmelzer, Atijeh Valai, and Michal Or-Guil. Affinity maturation of B cells involves not only a few but a whole spectrum of relevant mutations. *International Immunology*, 23(5):345–356, 2011.

[91] Adrian F. Ochsenbein, Daniel D. Pinschewer, Sophie Sierro, Edit Horvath, Hans Hengartner, and Rolf M. Zinkernagel. Protective long-term antibody memory by antigen-driven and T help-dependent differentiation of long-lived memory B cells to short-lived plasma cells independent of secondary lymphoid organs. *Proceedings of the National Academy of Sciences*, 97(24):13263 –13268, 2000.

[92] Edward S. Golub. Somatic mutation: Diversity and regulation of the immune repertoire. *Cell*, 48 (5):723–724, 1987.

[93] Joshua A. Weinstein, Ning Jiang, Richard A. White, Daniel S. Fisher, and Stephen R. Quake. High-Throughput Sequencing of the Zebrafish Antibody Repertoire. *Science*, 324(5928):807–810, 2009.

[94] Jennifer Benichou, Rotem Ben-Hamo, Yoram Louzoun, and Sol Efroni. Rep-Seq: uncovering the immunological repertoire through next-generation sequencing. *Immunology*, 135(3):183–191, 2012.

[95] Jay Shendure and Hanlee Ji. Next-generation DNA sequencing. *Nature Biotechnology*, 26(10): 1135–1145, 2008.

[96] Marcel Margulies, Michael Egholm, William E. Altman, Said Attiya, Joel S. Bader, Lisa A. Bemben, Jan Berka, Michael S. Braverman, Yi-Ju Chen, Zhoutao Chen, Scott B. Dewell, Lei Du, Joseph M. Fierro, Xavier V. Gomes, Brian C. Godwin, Wen He, Scott Helgesen, Chun He Ho, Gerard P. Irzyk, Szilveszter C. Jando, Maria L. I. Alenquer, Thomas P. Jarvie, Kshama B. Jirage, Jong-Bum Kim, James R. Knight, Janna R. Lanza, John H. Leamon, Steven M. Lefkowitz, Ming Lei, Jing Li, Kenton L. Lohman, Hong Lu, Vinod B. Makhijani, Keith E. McDade, Michael P. McKenna, Eugene W. Myers, Elizabeth Nickerson, John R. Nobile, Ramona Plant, Bernard P. Puc, Michael T. Ronan, George T. Roth, Gary J. Sarkis, Jan Fredrik Simons, John W. Simpson, Maithreyan Srinivasan, Karrie R. Tartaro, Alexander Tomasz, Kari A. Vogt, Greg A. Volkmer, Shally H. Wang, Yong Wang, Michael P. Weiner, Pengguang Yu, Richard F. Begley, and Jonathan M. Rothberg. Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, 437(7057):376–380, 2005.

[97] Thierry Mora, Aleksandra M. Walczak, William Bialek, and Curtis G. Callan. Maximum entropy models for antibody diversity. *Proceedings of the National Academy of Sciences*, 107(12):5405–5410, 2010.

[98] Ning Jiang, Joshua A. Weinstein, Lolita Penland, Richard A. White, Daniel S. Fisher, and Stephen R. Quake. Determinism and stochasticity during maturation of the zebrafish antibody repertoire. *Proceedings of the National Academy of Sciences*, 108(13):5348 –5353, 2011.

[99] Ponraj Prabakaran, Weizao Chen, Maria G Singarayan, Claudia C Stewart, Emily Streaker, Yang Feng, and Dimiter S Dimitrov. Expressed antibody repertoires in human cord blood cells: 454 sequencing and IMGT/HighV-QUEST analysis of germline gene usage, junctional diversity, and somatic mutations. *Immunogenetics*, 2011.

[100] Jacob Glanville, Wenwu Zhai, Jan Berka, Dilduz Telman, Gabriella Huerta, Gautam R. Mehta, Irene Ni, Li Mei, Purnima D. Sundar, Giles M. R. Day, David Cox, Arvind Rajpal, and Jaume Pons. Precise determination of the diversity of a combinatorial antibody library gives insight into the human immunoglobulin repertoire. *Proceedings of the National Academy of Sciences*, 106(48): 20216–20221, 2009.

[101] Ramy Arnaout, William Lee, Patrick Cahill, Tracey Honan, Todd Sparrow, Michael Weiand, Chad Nusbaum, Klaus Rajewsky, and Sergei B. Koralov. High-Resolution Description of Antibody Heavy-Chain Repertoires in Humans. *PLoS ONE*, 6(8):e22365, 2011.

[102] E. A Martin and Oxford University Press. *Concise medical dictionary.* Oxford University Press, Oxford; New York, 2007.

[103] Caroline Brissac, Alberto Nobrega, Jorge Carneiro, and John Stewart. Functional diversity of natural IgM. *Int. Immunol.*, 11(9):1501–1507, 1999.

[104] Victor Greiff, Henning Redestig, Juliane Luck, Nicole Bruni, Atijeh Valai, Susanne Hartmann, Sebastian Rausch, Johannes Schuchhardt, and Michal Or-Guil. A minimal model of peptide binding predicts ensemble properties of serum antibodies. *BMC Genomics*, 13(1):79, 2012.

[105] E. Helmreich, M. Kern, and H. N Eisen. The secretion of antibody by isolated lymph node cells. *Journal of Biological Chemistry*, 236(2):464, 1961.

[106] Toshifumi Hibi and Hans-Michael Dosch. Limiting dilution analysis of the B cell compartment in human bone marrow. *European Journal of Immunology*, 16(2):139–145, 1986.

[107] Sergio Arce, Elke Luger, Gwendolin Muehlinghaus, Giuliana Cassese, Anja Hauser, Alexander Horst, Katja Lehnert, Marcus Odendahl, Dirk Hönemann, Karl-Dieter Heller, Harald Kleinschmidt, Claudia Berek, Thomas Dörner, Veit Krenn, Falk Hiepe, Ralf Bargou, Andreas Radbruch, and Rudolf A Manz. CD38 Low IgG-Secreting Cells Are Precursors of Various CD38 High-Expressing Plasma Cell Populations. *Journal of Leukocyte Biology*, 75(6):1022–1028, 2004.

[108] J J Haaijman, H R Schuit, and W Hijmans. Immunoglobulin-containing cells in different lymphoid organs of the CBA mouse during its life-span. *Immunology*, 32(4):427–434, 1977.

[109] J A Brieva, E Roldán, M L De la Sen, and C Rodriguez. Human in vivo-induced spontaneous IgG-secreting cells from tonsil, blood and bone marrow exhibit different phenotype and functional level of maturation. *Immunology*, 72(4):580–583, 1991.

[110] Elizabeth J. Blink, Amanda Light, Axel Kallies, Stephen L. Nutt, Philip D. Hodgkin, and David M. Tarlinton. Early appearance of germinal center-derived memory B cells and plasma cells in blood after primary immunization. *J. Exp. Med.*, 201(4):545–554, 2005.

[111] P Vieira and K Rajewsky. The half-lives of serum immunoglobulins in adult mice. *European Journal of Immunology*, 18(2):313–316, 1988.

[112] Rudolf A. Manz, Anja E. Hauser, Falk Hiepe, and Andreas Radbruch. Maintenance of serum antibody levels. *Annual Review of Immunology*, 23(1):367–386, 2005.

[113] M A Kerr. The structure and function of human IgA. *Biochemical Journal*, 271(2):285–296, 1990.

[114] H Metzger. Structure and function of gamma M macroglobulins. *Advances in Immunology*, 12: 57–116, 1970.

[115] R. M. E. Parkhouse, Brigitte A. Askonas, and R. R. Dourmashkin. Electron microscopic studies of mouse immunoglobulin M; structure and reconstitution following reduction. *Immunology*, 18(4): 575–584, 1970.

[116] L A Herzenberg, A M Stall, P A Lalor, C Sidman, W A Moore, D R Parks, and L A Herzenberg. The Ly-1 B cell lineage. *Immunological Reviews*, 93:81–102, 1986.

[117] P Casali and E W Schettino. Structure and function of natural antibodies. *Current Topics in Microbiology and Immunology*, 210:167–179, 1996.

[118] Stephen V. Boyden, F.J. Dixon, and J.H. Humphrey. Natural Antibodies and the Immune Response. In *Advances in Immunology*, volume Volume 5, pages 1–28. Academic Press, 1966.

[119] Yang Yang, Eliver Eid Bou Ghosn, Leah E Cole, Tetyana V Obukhanych, Patricia Sadate-Ngatchou, Stefanie N Vogel, Leonard A Herzenberg, and Leonore A Herzenberg. Antigen-Specific Antibody Responses in B-1a and Their Relationship to Natural Immunity. *Proceedings of the National Academy of Sciences*, 109(14):5382–5387, 2012.

[120] F. Martin and J.F. Kearney. B-cell subsets and the mature preimmune repertoire. Marginal zone and B1 B cells as part of a "natural immune memory". *Immunological reviews*, 175(1):70–79, 2000.

[121] S. Bao, KW Beagley, AM Murray, V. Caristo, KI Matthaei, IG Young, and AJ Husband. Intestinal IgA plasma cells of the B1 lineage are IL-5 dependent. *Immunology*, 94(2):181–188, 1998.

[122] James W Tung, Matthew D Mrazek, Yang Yang, Leonard A Herzenberg, and Leonore A Herzenberg. Phenotypically Distinct B Cell Development Pathways Map to the Three B Cell Lineages in the Mouse. *Proceedings of the National Academy of Sciences*, 103(16):6293–6298, 2006.

[123] Atijeh Valai. *Migration and differentiation of murine germinal center derived B cell subsets in the course of the NP-specific immune response*. PhD thesis, Freie Universität Berlin, 2012.

[124] Christopher D.C. Allen, Takaharu Okada, and Jason G. Cyster. Germinal-Center Organization and Cellular Dynamics. *Immunity*, 27(2):190–202, 2007.

[125] Ian C. M. MacLennan, Kai-Michael Toellner, Adam F. Cunningham, Karine Serre, Daniel M.-Y. Sze, Elina Zuniga, Matthew C. Cook, and Carola G. Vinuesa. Extrafollicular antibody responses. *Immunological Reviews*, 194(1):8–18, 2003.

[126] H. P Roost, M. F Bachmann, A. Haag, U. Kalinke, V. Pliska, H. Hengartner, and R. M Zinkernagel. Early high-affinity neutralizing anti-viral IgG responses without further overall improvements of affinity. *Proceedings of the National Academy of Sciences of the United States of America*, 92(5): 1257, 1995.

[127] Rudolf A. Manz, Andreas Thiel, and Andreas Radbruch. Lifetime of plasma cells in the bone marrow. *Nature*, 388(6638):133–134, 1997.

[128] Mark K Slifka, Rustom Antia, Jason K Whitmire, and Rafi Ahmed. Humoral Immunity Due to Long-Lived Plasma Cells. *Immunity*, 8(3):363–372, 1998.

[129] Erika Hammarlund, Matthew W Lewis, Scott G Hansen, Lisa I Strelow, Jay A Nelson, Gary J Sexton, Jon M Hanifin, and Mark K Slifka. Duration of antiviral immunity after smallpox vaccination. *Nat Med*, 9(9):1131–1137, 2003.

[130] Ian J Amanna, Nichole E Carlson, and Mark K Slifka. Duration of humoral immunity to common viral and vaccine antigens. *The New England Journal of Medicine*, 357(19):1903–1915, 2007.

[131] Eric J Kunkel and Eugene C Butcher. Plasma-cell homing. *Nature Reviews. Immunology*, 3(10): 822–829, 2003.

[132] Stuart G. Tangye. Staying alive: regulation of plasma cell survival. *Trends in Immunology*, 32(12): 595–602, 2011.

[133] G Cassese, S Lindenau, B de Boer, S Arce, A Hauser, G Riemekasten, C Berek, F Hiepe, V Krenn, A Radbruch, and R A Manz. Inflamed kidneys of NZB / W mice are a major site for the homeostasis of plasma cells. *European Journal of Immunology*, 31(9):2726–2732, 2001.

[134] Taketoshi Yoshida, Henrik Mei, Thomas Dörner, Falk Hiepe, Andreas Radbruch, Simon Fillatreau, and Bimba F Hoyer. Memory B and memory plasma cells. *Immunological Reviews*, 237(1):117–139, 2010.

[135] M.F. Bachmann, T.M. Kundig, C.P. Kalberer, H. Hengartner, and R.M. Zinkernagel. How many specific B cells are needed to protect against a virus? *The Journal of Immunology*, 152(9):4235–4241, 1994.

[136] D Fleury, R S Daniels, J J Skehel, M Knossow, and T Bizebard. Structural evidence for recognition of a single epitope by two distinct antibodies. *Proteins*, 40(4):572–578, 2000.

[137] Andrew C.R. Martin, Janet C. Cheetham, Anthony R. Rees, and John J. Langone. [6] Molecular modeling of antibody combining sites. In *Molecular Design and Modeling: Concepts and Applications Part B: Antibodies and Antigens, Nucleic Acids, Polysaccharides, and Drugs*, volume Volume 203, pages 121–153. Academic Press, 1991.

[138] M K Gilson, J A Given, B L Bush, and J A McCammon. The statistical-thermodynamic basis for computation of binding affinities: a critical review. *Biophysical Journal*, 72(3):1047–1069, 1997.

[139] Stephanie Leavitt and Ernesto Freire. Direct measurement of protein binding energetics by isothermal titration calorimetry. *Current Opinion in Structural Biology*, 11(5):560–566, 2001.

[140] T Wiseman, S Williston, J F Brandts, and L N Lin. Rapid measurement of binding constants and heats of binding using a new titration calorimeter. *Analytical Biochemistry*, 179(1):131–137, 1989.

[141] J E Ladbury and B Z Chowdhry. Sensing the heat: the application of isothermal titration calorimetry to thermodynamic studies of biomolecular interactions. *Chemistry & Biology*, 3(10): 791–801, 1996.

[142] Stephen E. Harding and Babur Z. Chowdhry, editors. *Protein-ligand Interactions: Hydrodynamics and Calorimetry*. OUP Oxford, 2000.

[143] J. J. Christensen, R. M. Izatt, L. D. Hansen, and J. A. Partridge. Entropy Titration. A Calorimetric Method for the Determination of ΔG, ΔH, and ΔS from a Single Thermometric Titration. *J. Phys. Chem.*, 70(6):2003–2010, 1966.

[144] J.Doyne Farmer, Norman H Packard, and Alan S Perelson. The immune system, adaptation, and machine learning. *Physica D: Nonlinear Phenomena*, 22(1–3):187–204, 1986.

[145] R. Hightower, S. Forrest, and A.S. Perelson. The Evolution of Cooperation in Immune System Gene Libraries l. *Working Paper*, 1992.

[146] D Lancet, E Sadovsky, and E Seidemann. Probability model for molecular recognition in biological receptor repertoires: significance to the olfactory system. *Proceedings of the National Academy of Sciences of the United States of America*, 90(8):3715–3719, 1993.

[147] J D Farmer, S A Kauffman, N H Packard, and A S Perelson. Adaptive dynamic networks as models for the immune system and autocatalytic sets. *Annals of the New York Academy of Sciences*, 504: 118–131, 1987.

[148] Frederic A. Fellouse, Bing Li, Deanne M. Compaan, Andrew A. Peden, Sarah G. Hymowitz, and Sachdev S. Sidhu. Molecular Recognition by a Binary Code. *Journal of Molecular Biology*, 348(5): 1153–1162, 2005.

[149] A. Sircar, E. T. Kim, and J. J. Gray. RosettaAntibody: antibody variable region homology modeling server. *Nucleic Acids Research*, 37(Web Server):W474–W479, 2009.

[150] Louis A. Clark, P. Ann Boriack-Sjodin, John Eldredge, Christopher Fitch, Bethany Friedman, Karl J.M. Hanf, Matthew Jarpe, Stefano F. Liparoto, You Li, Alexey Lugovskoy, Stephan Miller, Mia Rushe, Woody Sherman, Kenneth Simon, and Herman Van Vlijmen. Affinity enhancement of an in vivo matured therapeutic antibody using structure-based computational design. *Protein Science : A Publication of the Protein Society*, 15(5):949–960, 2006.

[151] Annemarie Honegger, Alain Daniel Malebranche, Daniela Röthlisberger, and Andreas Plückthun. The Influence of the Framework Core Residues on the Biophysical Properties of Immunoglobulin Heavy Chain Variable Domains. *Protein Engineering Design and Selection*, 22(3):121–134, 2009.

[152] Peter T. Jones, Paul H. Dear, Jefferson Foote, Michael S. Neuberger, and Greg Winter. Replacing the complementarity-determining regions in a human antibody with those from a mouse. *, Published online: 29 May 1986; | doi:10.1038/321522a0*, 321(6069):522–525, 1986.

[153] Martin Schlapschy, Marton Fogarasi, Helga Gruber, Oliver Gresch, Claudia Schäfer, Yasmine Aguib, and Arne Skerra. Functional Humanization of an Anti-CD16 Fab Fragment: Obstacles of Switching from Murine λ to Human λ or κ Light Chains. *Protein Engineering Design and Selection*, 22(3):175–188, 2009.

[154] Shuchismita Dutta, Kyle Burkhardt, Jasmine Young, Ganesh J Swaminathan, Takanori Matsuura, Kim Henrick, Haruki Nakamura, and Helen M Berman. Data deposition and annotation at the worldwide protein data bank. *Molecular Biotechnology*, 42(1):1–13, 2009.

[155] Arvind Sivasubramanian, Aroop Sircar, Sidhartha Chaudhury, and Jeffrey J Gray. Toward high-resolution homology modeling of antibody Fv regions and application to antibody-antigen docking. *Proteins: Structure, Function, and Bioinformatics*, 74(2):497–514, 2009.

[156] D.M. Webster and A.R. Rees. Molecular modeling of antibody-combining sites. *Methods Mol Biol*, 51:17–49, 1995.

[157] D.M. Webster, J. Pedersen, D. Staunton, A. Jones, and A.R. Rees. Antibody-combining sites. *Applied Biochemistry and Biotechnology*, 47(2):119–134, 1994.

[158] D R Davies and H Metzger. Structural Basis of Antibody Function. *Annual Review of Immunology*, 1(1):87–115, 1983.

[159] Gary Walsh. Biopharmaceutical benchmarks 2006. *Nature Biotechnology*, 24(7):769–776, 2006.

[160] Janice Reichert and Alex Pavlou. Monoclonal antibodies market. *Nature Reviews Drug Discovery*, 3(5):383–384, 2004.

[161] C Chothia and A M Lesk. Canonical structures for the hypervariable regions of immunoglobulins. *Journal of Molecular Biology*, 196(4):901–917, 1987.

[162] B Al-Lazikani, A M Lesk, and C Chothia. Standard conformations for the canonical structures of immunoglobulins. *Journal of Molecular Biology*, 273(4):927–948, 1997.

[163] Veronica Morea, Arthur M. Lesk, and Anna Tramontano. Antibody Modeling: Implications for Engineering and Design. *Methods*, 20(3):267–279, 2000.

[164] Benjamin North, Andreas Lehmann, and Roland L. Dunbrack Jr. A New Clustering of Antibody CDR Loop Conformations. *Journal of Molecular Biology*, 406(2):228–256, 2011.

[165] Yoonjoo Choi and Charlotte M Deane. Predicting antibody complementarity determining region structures without classification. *Molecular bioSystems*, 7(12):3327–3334, 2011.

[166] N. Whitelegg and A.R. Rees. Antibody Variable Regions. *Antibody engineering: methods and protocols*, 248:51, 2004.

[167] Paolo Marcatili, Alessandra Rosi, and Anna Tramontano. PIGS: Automatic Prediction of Antibody Structures. *Bioinformatics*, 24(17):1953–1954, 2008.

[168] Gregory L. Warren, C. Webster Andrews, Anna-Maria Capelli, Brian Clarke, Judith LaLonde, Millard H. Lambert, Mika Lindvall, Neysa Nevins, Simon F. Semus, Stefan Senger, Giovanna Tedesco, Ian D. Wall, James M. Woolven, Catherine E. Peishoff, and Martha S. Head. A Critical Assessment of Docking Programs and Scoring Functions. *J. Med. Chem.*, 49(20):5912–5931, 2005.

[169] S. Wold. Chemometrics; what do we mean with it, and what do we want from it? *Chemometrics and Intelligent Laboratory Systems*, 30(1):109–115, 1995.

[170] Salvador Eugenio C. Caoili. Benchmarking B-Cell Epitope Prediction for the Design of Peptide-Based Vaccines: Problems and Prospects. *Journal of Biomedicine and Biotechnology*, 2010:1–15, 2010.

[171] Marc H. V. Van Van Regenmortel. The rational design of biological complexity: A deceptive metaphor. *PROTEOMICS*, 7(6):965–975, 2007.

[172] Martin J. Blythe and Darren R. Flower. Benchmarking B cell epitope prediction: Underperformance of existing methods. *Protein Science : A Publication of the Protein Society*, 14(1):246–248, 2005.

[173] Galina F Denisova, Dimitri A Denisov, and Jonathan L Bramson. Applying bioinformatics for antibody epitope prediction using affinity-selected mimotopes – relevance for vaccine design. *Immunome Research*, 6(Suppl 2):S6, 2010.

[174] Virginie Lollier, Sandra Denery-Papini, Colette Larré, and Dominique Tessier. A generic approach to evaluate how B-cell epitopes are surface-exposed on protein structures. *Molecular Immunology*, 48(4):577–585, 2011.

[175] Lilian Lacerda Bueno, Francisco Pereira Lobo, Cristiane Guimarães Morais, Luíza Carvalho Mourão, Ricardo Andrez Machado de Ávila, Irene Silva Soares, Cor Jesus Fontes, Marcus Vinícius Lacerda, Carlos Chavez Olórtegui, Daniella Castanheira Bartholomeu, Ricardo Toshio Fujiwara, and Érika Martins Braga. Identification of a Highly Antigenic Linear B Cell Epitope within Plasmodium vivax Apical Membrane Antigen 1 (AMA-1). *PLoS ONE*, 6(6):e21289, 2011.

[176] Darren R. Flower. *Immunoinformatics: Predicting Immunogenicity In Silico*. Humana Press, 1 edition, 2007.

[177] Yasser EL-Manzalawy and Vasant Honavar. Recent advances in B-cell epitope prediction methods. *Immunome Research*, 6(Suppl 2):S2, 2010.

[178] I.M. Roitt and P.J. Delves. *Roitt's Essential Immunology*. Essentials Series. Blackwell Science, 2001. ISBN 9780632059027. URL `http://books.google.de/books?id=SEAVHm1QQQAC`.

[179] Julia Ponomarenko, Nikitas Papangelopoulos, Dirk M. Zajonc, Bjoern Peters, Alessandro Sette, and Philip E. Bourne. IEDB-3D: structural data within the immune epitope database. *Nucleic Acids Research*, 39(Database issue):D1164–D1170, 2011.

[180] J. K. Scott and G. P. Smith. Searching for Peptide Ligands with an Epitope Library. *Science*, 249 (4967):386–390, 1990.

[181] R H Meloen, W C Puijk, and J W Slootstra. Mimotopes: realization of an unlikely concept. *Journal of Molecular Recognition: JMR*, 13(6):352–359, 2000.

[182] Jason A Greenbaum, Pernille Haste Andersen, Martin Blythe, Huynh-Hoa Bui, Raul E Cachau, James Crowe, Matthew Davies, A. S Kolaskar, Ole Lund, Sherrie Morrison, Brendan Mumey, Yanay Ofran, Jean-Luc Pellequer, Clemencia Pinilla, Julia V Ponomarenko, G. P. S Raghava, Marc H. V van Regenmortel, Erwin L Roggen, Alessandro Sette, Avner Schlessinger, Johannes Sollner, Martin Zand, and Bjoern Peters. Towards a consensus on datasets and evaluation metrics for developing B-cell epitope prediction tools. *Journal of Molecular Recognition*, 20(2):75–82, 2007.

[183] A Zvirbliene, I Kucinskaite, I Sezaite, D Samuel, and K Sasnauskas. Mapping of B cell epitopes in measles virus nucleocapsid protein. *Archives of Virology*, 152(1):25–39, 2007.

[184] M Fernandez-Alonso, G Lorenzo, L Perez, R Bullido, A Estepa, N Lorenzen, and J M Coll. Mapping of linear antibody epitopes of the glycoprotein of VHSV, a salmonid rhabdovirus. *Diseases of Aquatic Organisms*, 34(3):167–176, 1998.

[185] John Mark Carter and Larry Loomis-Price. B cell epitope mapping using synthetic peptides. *Current Protocols in Immunology / Edited by John E. Coligan ... [et Al.]*, Chapter 9:Unit 9.4, 2004.

[186] J M Carter. Epitope mapping of a protein using the Geysen (PEPSCAN) procedure. *Methods in Molecular Biology (Clifton, N.J.)*, 36:207–223, 1994.

[187] Achim Kramer and Jens Schneider-Mergener. Synthesis and Screening of Peptide Libraries on Continuous Cellulose Membrane Supports. In Shmuel Cabilly, editor, *Combinatorial Peptide Library Protocols*, volume 87 of *Methods in Molecular Biology*, pages 25–39. Humana Press, 1998.

[188] Juliane Bongartz, Nicole Bruni, and Michal Or-Guil. Epitope mapping using randomly generated peptide libraries. *Methods in Molecular Biology (Clifton, N.J.)*, 524:237–246, 2009.

[189] Joseph Barten Legutki, D Mitchell Magee, Phillip Stafford, and Stephen Albert Johnston. A general method for characterization of humoral immunity induced by a vaccine or infection. *Vaccine*, 28 (28):4529–4537, 2010.

[190] Ulrich Reineke and Robert Sabat. Antibody epitope mapping using SPOT peptide arrays. *Methods in Molecular Biology (Clifton, N.J.)*, 524:145–167, 2009.

[191] M Sandberg, L Eriksson, J Jonsson, M Sjöström, and S Wold. New chemical descriptors relevant for the design of biologically active peptides. A multivariate characterization of 87 amino acids. *Journal of Medicinal Chemistry*, 41(14):2481–2491, 1998.

[192] T P Hopp and K R Woods. Prediction of protein antigenic determinants from amino acid sequences. *Proceedings of the National Academy of Sciences of the United States of America*, 78(6):3824–3828, 1981.

[193] M Levitt. A simplified representation of protein conformations for rapid simulation of protein folding. *Journal of Molecular Biology*, 104(1):59–107, 1976.

[194] J M Parker, D Guo, and R S Hodges. New hydrophilicity scale derived from high-performance liquid chromatography peptide retention data: correlation of predicted surface residues with antigenicity and X-ray-derived accessible sites. *Biochemistry*, 25(19):5425–5432, 1986.

[195] P. A. Karplus and G. E. Schulz. Prediction of chain flexibility in proteins. *Naturwissenschaften*, 72 (4):212–213, 1985.

[196] J.L. Pellequer, E. Westhof, M.H.V. Van Regenmortel, and John J. Langone. [8] Predicting location of continuous epitopes in proteins from their primary structures. In *Molecular Design and Modeling: Concepts and Applications Part B: Antibodies and Antigens, Nucleic Acids, Polysaccharides, and Drugs*, volume Volume 203, pages 176–201. Academic Press, 1991.

[197] E A Emini, J V Hughes, D S Perlow, and J Boger. Induction of hepatitis A virus-neutralizing antibody by a virus-specific synthetic peptide. *Journal of Virology*, 55(3):836–839, 1985.

[198] J L Pellequer, E Westhof, and M H Van Regenmortel. Correlation between the location of antigenic sites and the prediction of turns in proteins. *Immunology Letters*, 36(1):83–99, 1993.

[199] A J Alix. Predictive estimation of protein linear epitopes by using the program PEOPLE. *Vaccine*, 18(3-4):311–314, 1999.

[200] Michael Odorico and Jean-Luc Pellequer. BEPITOPE: predicting the location of continuous epitopes and patterns in proteins. *Journal of Molecular Recognition: JMR*, 16(1):20–22, 2003.

[201] J Chen, H Liu, J Yang, and K-C Chou. Prediction of linear B-cell epitopes using amino acid pair antigenicity scale. *Amino Acids*, 33(3):423–428, 2007.

[202] Jens Erik Pontoppidan Larsen, Ole Lund, and Morten Nielsen. Improved method for predicting linear B-cell epitopes. *Immunome Research*, 2:2, 2006.

[203] Randi Vita, Laura Zarebski, Jason A. Greenbaum, Hussein Emami, Ilka Hoof, Nima Salimi, Rohini Damle, Alessandro Sette, and Bjoern Peters. The Immune Epitope Database 2.0. *Nucleic Acids Research*, 38(Database issue):D854–D862, 2010.

[204] Johannes Söllner and Bernd Mayer. Machine learning approaches for prediction of linear B-cell epitopes on proteins. *Journal of Molecular Recognition*, 19(3):200–208, 2006.

[205] Sudipto Saha and G P S Raghava. Prediction of continuous B-cell epitopes in an antigen using recurrent neural network. *Proteins*, 65(1):40–48, 2006.

[206] Michael J Sweredoski and Pierre Baldi. COBEpro: a novel system for predicting continuous B-cell epitopes. *Protein Engineering, Design & Selection: PEDS*, 22(3):113–120, 2009.

[207] Y Feng, F Jacobs, E Van Craeyveld, J Lievens, J Snoeys, S Van Linthout, and B De Geest. The impact of antigen expression in antigen-presenting cells on humoral immune responses against the transgene product. *Gene Therapy*, 17(2):288–293, 2010.

[208] Arthur M. Silverstein. *A History of Immunology, Second Edition*. Academic Press, 2 edition, 2009.

[209] F JFrancisco J. Quintana, Y Merbl, E Sahar, E Domany, and I R Cohen. Antigen-chip technology for accessing global information about the state of the body. *Lupus*, 15(7):428–430, 2006.

[210] Phillip Stafford, Rebecca Halperin, Joseph Bart Legutki, Dewey Mitchell Magee, John Galgiani, and Stephen Albert Johnston. Physical Characterization of the "Immunosignaturing Effect". *Molecular & Cellular Proteomics*, 11(4), 2012.

[211] A Nobrega, M Haury, A Grandien, E Malanchère, A Sundblad, and A Coutinho. Global analysis of antibody repertoires. II. Evidence for specificity, self-selection and the immunological "homunculus" of antibodies in normal serum. *European Journal of Immunology*, 23(11):2851–2859, 1993.

[212] M Haury, A Grandien, A Sundblad, A Coutinho, and A Nobrega. Global analysis of antibody repertoires. 1. An immunoblot method for the quantitative screening of a large number of reactivities. *Scandinavian Journal of Immunology*, 39(1):79–87, 1994.

[213] Jens Rauch and Olivier Gires. SEREX, Proteomex, AMIDA, and beyond: Serological screening technologies for target identification. *PROTEOMICS - CLINICAL APPLICATIONS*, 2(3):355–371, 2008.

[214] Lina Cekaite, Ola Haug, Ola Myklebost, Magne Aldrin, Bjørn Østenstad, Marit Holden, Arnoldo Frigessi, Eivind Hovig, and Mouldy Sioud. Analysis of the humoral immune response to immunoselected phage-displayed peptides by a microarray-based method. *PROTEOMICS*, 4(9):2572–2582, 2004.

[215] Paul J. Mintz, Jeri Kim, Kim-Anh Do, Xuemei Wang, Ralph G. Zinner, Massimo Cristofanilli, Marco A. Arap, Waun Ki Hong, Patricia Troncoso, Christopher J. Logothetis, Renata Pasqualini, and Wadih Arap. Fingerprinting the circulating repertoire of antibodies from cancer patients. *Nat Biotech*, 21(1):57–63, 2003.

[216] Madhumita Chatterjee, Saroj Mohapatra, Alexei Ionan, Gagandeep Bawa, Rouba Ali-Fehmi, Xiaoju Wang, James Nowak, Bin Ye, Fatimah A. Nahhas, Karen Lu, Steven S. Witkin, David Fishman, Adnan Munkarah, Robert Morris, Nancy K. Levin, Natalie N. Shirley, Gerard Tromp, Judith Abrams, Sorin Draghici, and Michael A. Tainsky. Diagnostic Markers of Ovarian Cancer by High-Throughput Antigen Cloning and Detection on Arrays. *Cancer Res*, 66(2):1181–1190, 2006.

[217] Atsushi Kuno, Noboru Uchiyama, Shiori Koseki-Kuno, Youji Ebe, Seigo Takashima, Masao Yamada, and Jun Hirabayashi. Evanescent-field fluorescence-assisted lectin microarray: a new strategy for glycan profiling. *Nature Methods*, 2(11):851–856, 2005.

[218] Oyindasola Oyelaran and Jeffrey C Gildersleeve. Glycan arrays: recent advances and future challenges. *Current Opinion in Chemical Biology*, 13(4):406–413, 2009.

[219] Henning Ulrich and Carsten Wrenger. Disease-specific biomarker discovery by aptamers. *Cytometry Part A*, 75A(9):727–733, 2009.

[220] Kyung-Mi Song, Seonghwan Lee, and Changill Ban. Aptamers and Their Biological Applications. *Sensors (Basel, Switzerland)*, 12(1):612–631, 2012.

[221] Andrew D. Ellington and Jack W. Szostak. In vitro selection of RNA molecules that bind specific ligands. *, Published online: 30 August 1990; | doi:10.1038/346818a0*, 346(6287):818–822, 1990.

[222] Asaf Madi, Inbal Hecht, Sharron Bransburg-Zabary, Yifat Merbl, Adi Pick, Merav Zucker-Toledano, Francisco J Quintana, Alfred I Tauber, Irun R Cohen, and Eshel Ben-Jacob. Organization of the autoantibody repertoire in healthy newborns and adults revealed by system level informatics of antigen microarray data. *Proceedings of the National Academy of Sciences of the United States of America*, 106(34):14484–14489, 2009.

[223] Heiko Andresen and Carsten Grotzinger. Deciphering the Antibodyome - Peptide Arrays for Serum Antibody Biomarker Diagnostics. *Current Proteomics*, 6:1–12, 2009.

[224] Yifat Merbl, Royi Itzchak, Tal Vider-Shalit, Yoram Louzoun, Francisco J. Quintana, Ezra Vadai, Lea Eisenbach, and Irun R Cohen. A systems immunology approach to the host-tumor interaction: large-scale patterns of natural autoantibodies distinguish healthy and tumor-bearing mice. *PloS One*, 4(6):e6053, 2009.

[225] Yu M Foong, Jiaqi Fu, Shao Q Yao, and Mahesh Uttamchandani. Current advances in peptide and small molecule microarray technologies. *Current Opinion in Chemical Biology*, 16(1–2):234–242, 2012.

[226] Stephen F. Kingsmore. Multiplexed protein measurement: technologies and applications of protein and antibody arrays. *Nature Reviews Drug Discovery*, 5(4):310–321, 2006.

[227] Marie-Laure Lesaicherre, Mahesh Uttamchandani, Grace Y.J Chen, and Shao Q Yao. Developing site-Specific immobilization strategies of peptides in a microarray. *Bioorganic & Medicinal Chemistry Letters*, 12(16):2079–2083, 2002.

[228] Gavin MacBeath and Stuart L. Schreiber. Printing Proteins as Microarrays for High-Throughput Function Determination. *Science*, 289(5485):1760 –1763, 2000.

[229] Tadashi Okamoto, Tomohiro Suzuki, and Nobuko Yamamoto. Microarray fabrication with covalent attachment of DNA using Bubble Jet technology. *Nature Biotechnology*, 18(4):438–441, 2000.

[230] Virginia Espina, Elisa C. Woodhouse, Julia Wulfkuhle, Heather D. Asmussen, Emanuel F. Petricoin III, and Lance A. Liotta. Protein microarray detection strategies: focus on direct detection technologies. *Journal of Immunological Methods*, 290(1–2):121–133, 2004.

[231] Armin A. Weiser, Michal Or-Guil, Victor Tapia, Astrid Leichsenring, Johannes Schuchhardt, Cornelius Frömmel, and Rudolf Volkmer-Engert. SPOT synthesis: Reliability of array-based measurement of peptide binding affinity. *Analytical Biochemistry*, 342(2):300–311, 2005.

[232] Victor Tapia, Juliane Bongartz, Mike Schutkowski, Nicole Bruni, Armin Weiser, Bernhard Ay, Rudolf Volkmer, and Michal Or-Guil. Affinity profiling using the peptide microarray technology: A case study. *Analytical Biochemistry*, 363(1):108–118, 2007.

[233] Kaitlin Kroening, Stephen Albert Johnston, and Joseph Barten Legutki. Autoreactive antibodies raised by self derived de novo peptides can identify unrelated antigens on protein microarrays. Are autoantibodies really autoantibodies? *Experimental and molecular pathology*, 92(3):304–311, 2012.

[234] Wolfgang Hueber, Brian A. Kidd, Beren H. Tomooka, Byung J. Lee, Bonnie Bruce, James F. Fries, Grete Sønderstrup, Paul Monach, Jan W. Drijfhout, Walther J. van Venrooij, Paul J. Utz, Mark C. Genovese, and William H. Robinson. Antigen microarray profiling of autoantibodies in rheumatoid arthritis. *Arthritis & Rheumatism*, 52(9):2645–2655, 2005.

[235] William H. Robinson, Carla DiGennaro, Wolfgang Hueber, Brian B. Haab, Makoto Kamachi, Erik J. Dean, Sylvie Fournel, Derek Fong, Mark C. Genovese, Henry E. Neuman de Vegvar, Karl Skriner, David L. Hirschberg, Robert I. Morris, Sylviane Muller, Ger J. Pruijn, Walther J. van Venrooij, Josef S. Smolen, Patrick O. Brown, Lawrence Steinman, and Paul J. Utz. Autoantigen microarrays for multiplex characterization of autoantibody responses. *Nat Med*, 8(3):295–301, 2002.

[236] William H Robinson, Paulo Fontoura, Byung J Lee, Henry E Neuman de Vegvar, Jennifer Tom, Rosetta Pedotti, Carla D DiGennaro, Dennis J Mitchell, Derek Fong, Peggy P-K Ho, Pedro J Ruiz, Emanual Maverakis, David B Stevens, Claude C A Bernard, Roland Martin, Vijay K Kuchroo, Johannes M van Noort, Claude P Genain, Sandra Amor, Tomas Olsson, Paul J Utz, Hideki Garren, and Lawrence Steinman. Protein microarrays guide tolerizing DNA vaccine treatment of autoimmune encephalomyelitis. *Nat Biotech*, 21(9):1033–1039, 2003.

[237] Simani Gaseitsiwe, Davide Valentini, Shahnaz Mahdavifar, Isabelle Magalhaes, Daniel F. Hoft, Johannes Zerweck, Mike Schutkowski, Jan Andersson, Marie Reilly, and Markus J. Maeurer. Pattern Recognition in Pulmonary Tuberculosis Defined by High Content Peptide Microarray Chip Analysis Representing 61 Proteins from M. tuberculosis. *PLoS ONE*, 3(12):e3840, 2008.

[238] Michael Hecker, Peter Lorenz, Felix Steinbeck, Li Hong, Gabriela Riemekasten, Yixue Li, Uwe K. Zettl, and Hans-Jürgen Thiesen. Computational analysis of high-density peptide microarray data with application from systemic sclerosis to multiple sclerosis. *Autoimmunity Reviews*, 11(3):180–190, 2012.

[239] M. Muralidhar Reddy, Rosemary Wilson, Johnnie Wilson, Steven Connell, Anne Gocke, Linda Hynan, Dwight German, and Thomas Kodadek. Identification of Candidate IgG Biomarkers for Alzheimer's Disease via Combinatorial Library Screening. *Cell*, 144:132–142, 2011.

[240] Audrey Kauffmann, Robert Gentleman, and Wolfgang Huber. arrayQualityMetrics–a bioconductor package for quality assessment of microarray data. *Bioinformatics*, 25(3):415–416, 2009.

[241] Jeremy Gollub, Catherine A Ball, Gail Binkley, Janos Demeter, David B Finkelstein, Joan M Hebert, Tina Hernandez-Boussard, Heng Jin, Miroslava Kaloper, John C Matese, Mark Schroeder, Patrick O Brown, David Botstein, and Gavin Sherlock. The Stanford Microarray Database: Data Access and Quality Assessment Tools. *Nucleic Acids Research*, 31(1):94–96, 2003.

[242] John Quackenbush. Microarray data normalization and transformation. *Nature Genetics*, 32(Supp): 496–501, 2002.

[243] George C Tseng, Min-Kyu Oh, Lars Rohlin, James C Liao, and Wing Hung Wong. Issues in cDNA Microarray Analysis: Quality Filtering, Channel Normalization, Models of Variations and Assessment of Gene Effects. *Nucleic Acids Research*, 29(12):2549–2557, 2001.

[244] Leming Shi, Laura H Reid, Wendell D Jones, Richard Shippy, Janet A Warrington, Shawn C Baker, Patrick J Collins, Francoise de Longueville, Ernest S Kawasaki, Kathleen Y Lee, Yuling Luo, Yongming Andrew Sun, James C Willey, Robert A Setterquist, Gavin M Fischer, Weida Tong, Yvonne P Dragan, David J Dix, Felix W Frueh, Frederico M Goodsaid, Damir Herman, Roderick V Jensen, Charles D Johnson, Edward K Lobenhofer, Raj K Puri, Uwe Schrf, Jean Thierry-Mieg, Charles Wang, Mike Wilson, Paul K Wolber, Lu Zhang, Shashi Amur, Wenjun Bao, Catalin C Barbacioru, Anne Bergstrom Lucas, Vincent Bertholet, Cecilie Boysen, Bud Bromley, Donna Brown, Alan Brunner, Roger Canales, Xiaoxi Megan Cao, Thomas A Cebula, James J Chen, Jing Cheng, Tzu-Ming Chu, Eugene Chudin, John Corson, J Christopher Corton, Lisa J Croner, Christopher Davies, Timothy S Davison, Glenda Delenstarr, Xutao Deng, David Dorris, Aron C Eklund, Xiao-Hui Fan, Hong Fang, Stephanie Fulmer-Smentek, James C Fuscoe, Kathryn Gallagher, Weigong Ge, Lei Guo, Xu Guo, Janet Hager, Paul K Haje, Jing Han, Tao Han, Heather C Harbottle, Stephen C Harris, Eli Hatchwell, Craig A Hauser, Susan Hester, Huixiao Hong, Patrick Hurban, Scott A Jackson, Hanlee Ji, Charles R Knight, Winston P Kuo, J Eugene Leclerc, Shawn Levy, Quan-Zhen Li, Chunmei Liu, Ying Liu, Michael J Lombardi, Yunqing Ma, Scott R Magnuson, Botoul Maqsodi, Tim McDaniel, Nan Mei, Ola Myklebost, Baitang Ning, Natalia Novoradovskaya, Michael S Orr, Terry W Osborn, Adam Papallo, Tucker A Patterson, Roger G Perkins, Elizabeth H Peters, Ron Peterson, Kenneth L Philips, P Scott Pine, Lajos Pusztai, Feng Qian, Hongzu Ren, Mitch Rosen, Barry A Rosenzweig, Raymond R Samaha, Mark Schena, Gary P Schroth, Svetlana Shchegrova, Dave D Smith, Frank Staedtler, Zhenqiang Su, Hongmei Sun, Zoltan Szallasi, Zivana Tezak, Danielle Thierry-Mieg, Karol L Thompson, Irina Tikhonova, Yaron Turpaz, Beena Vallanat, Christophe Van, Stephen J Walker, Sue Jane Wang, Yonghong Wang, Russ Wolfinger, Alex Wong, Jie Wu, Chunlin Xiao, Qian Xie, Jun Xu, Wen Yang, Liang Zhang, Sheng Zhong, Yaping Zong, and William Slikker. The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. *Nature Biotechnology*, 24(9):1151–1161, 2006.

[245] Justin R Brown, Phillip Stafford, Stephen A Johnston, and Valentin Dinu. Statistical methods for analyzing immunosignatures. *BMC Bioinformatics*, 12:349, 2011.

[246] M. Kukreja, S.A. Johnston, and P. Stafford. Comparative study of classification algorithms for immunosignaturing data. *BMC bioinformatics*, 13(1):139, 2012.

[247] T. Nahtman, A. Jernberg, S. Mahdavifar, J. Zerweck, M. Schutkowski, M. Maeurer, and M. Reilly. Validation of peptide epitope microarray experiments and extraction of quality data. *Journal of Immunological Methods*, 328(1-2):1–13, 2007.

[248] Marie Reilly and Davide Valentini. Visualisation and pre-processing of peptide microarray data. *Methods in Molecular Biology (Clifton, N.J.)*, 570:373–389, 2009.

[249] Bernhard Renard, Martin Lower, Yvonne Kuhne, Ulf Reimer, Andree Rothermel, Ozlem Tureci, John Castle, and Ugur Sahin. rapmad: Robust analysis of peptide microarray data. *BMC Bioinformatics*, 12(1):324, 2011.

[250] Shu-Wen W Chen, Marc H V Van Regenmortel, and Jean-Luc Pellequer. Structure-activity relationships in peptide-antibody complexes: implications for epitope prediction and development of synthetic peptide vaccines. *Current Medicinal Chemistry*, 16(8):953–964, 2009.

[251] Fabien Pamelard, Gael Even, Costin Apostol, Cristian Preda, Clarisse Dhaenens, Vronique Fafeur, Rémi Desmet, and Oleg Melnyk. PASE: A Web-Based Platform for Peptide/Protein Microarray Experiments. *Methods Mol Biol.*, 570:413–430, 2009.

[252] J M Behnke, D M Menge, and H Noyes. Heligmosomoides bakeri: a model for exploring the biology and genetics of resistance to chronic gastrointestinal nematode infections. *Parasitology*, 136(12): 1565–1580, 2009.

[253] Jerzy Behnke and Phil D Harris. Heligmosomoides bakeri: a new name for an old worm? *Trends in Parasitology*, 2010.

[254] F G Monroy and F J Enriquez. Heligmosomoides polygyrus: a model for chronic gastrointestinal helminthiasis. *Parasitology Today (Personal Ed.)*, 8(2):49–54, 1992.

[255] Robert M Anthony, Joseph F Urban, Farhang Alem, Hossein A Hamed, Cristina T Rozo, Jean-Luc Boucher, Nico Van Rooijen, and William C Gause. Memory TH2 cells induce alternatively activated macrophages to mediate protection against nematode parasites. *Nature medicine*, 12(8):955–960, 2006.

[256] Kathy D. McCoy, Maaike Stoel, Rebecca Stettler, Patrick Merky, Katja Fink, Beatrice M. Senn, Corinne Schaer, Joanna Massacand, Bernhard Odermatt, Hans C. Oettgen, Rolf M. Zinkernagel, Nicolaas A. Bos, Hans Hengartner, Andrew J. Macpherson, and Nicola L. Harris. Polyclonal and Specific Antibodies Mediate Protective Immunity against Enteric Helminth Infection. *Cell Host & Microbe*, 4(4):362–373, 2008.

[257] Nicola Harris and William C. Gause. To B or not to B: B cells and the Th2-type immune response to helminths. *Trends in Immunology*, 32(2):80–88, 2011.

[258] Wojciech Wojciechowski, David P. Harris, Frank Sprague, Betty Mousseau, Melissa Makris, Kim Kusser, Tasuko Honjo, Katja Mohrs, Markus Mohrs, and Troy Randall. Cytokine-Producing Effector B Cells Regulate Type 2 Immunity to H. polygyrus. *Immunity*, 30(3):421–433, 2009.

[259] Juliane Lück. *Technologische Bewertung von Peptid-Mikroarrays als Methode der serologischen Diagnostik*. Biowissenschaften, Biologie, Humboldt-Universität zu Berlin, Berlin, 2012.

[260] Kristina Gruden, Matjaž Hren, Ana Herman, Andrej Blejec, Tanja Albrecht, Joachim Selbig, Chris Bauer, Johannes Schuchardt, Michal Or-Guil, Klemen Zupančič, Urban Svajger, Borut Stabuc, Alojz Ihan, Andreja Nataša Kopitar, Maja Ravnikar, Miomir Knežević, Primož Rožman, and Matjaž Jeras. A "crossomics" study analysing variability of different components in peripheral blood of healthy caucasoid individuals. *PloS One*, 7(1):e28761, 2012.

[261] K. Solez, R. B Colvin, L. C Racusen, M. Haas, B. Sis, M. Mengel, P. F Halloran, W. Baldwin, G. Banfi, A. B Collins, F. Cosio, D. S. R David, C. Drachenberg, G. Einecke, A. B Fogo, I. W Gibson, D. Glotz, S. S Iskandar, E. Kraus, E. Lerut, R. B Mannon, M. Mihatsch, B. J Nankivell, V. Nickeleit, J. C Papadimitriou, P. Randhawa, H. Regele, K. Renaudin, I. Roberts, D. Seron, R. N Smith, and M. Valente. Banff 07 Classification of Renal Allograft Pathology: Updates and Future Directions. *American Journal of Transplantation*, 8(4):753–760, 2008.

[262] Lorraine C Racusen, Kim Solez, Robert B Colvin, Stephen M Bonsib, Maria C Castro, Tito Cavallo, Byron P Croker, A Jake Demetris, Cynthia B Drachenberg, Agnes B Fogo, Peter Furness, Lillian W Gaber, Ian W Gibson, Dennis Glotz, Julio C Goldberg, Joseph Grande, Philip F Halloran, H E Hansen, Barry Hartley, Pekka J Hayry, Claire M Hill, Ernesto O Hoffman, Lawrence G Hunsicker, Anne S Lindblad, Niels Marcussen, Michael J Mihatsch, Tibor Nadasdy, Peter Nickerson, T Steen Olsen, John C Papadimitriou, Parmjeet S Randhawa, David C Rayner, Ian Roberts, Stephen Rose, David Rush, Luis Salinas-Madrigal, Daniel R Salomon, Stale Sund, Eero Taskinen, Kiril Trpkov, and Yutaka Yamaguchi. The Banff 97 working classification of renal allograft pathology. *Kidney International*, 55(2):713–723, 1999.

[263] Eric Meffre, Anne Schaefer, Hedda Wardemann, Patrick Wilson, Eric Davis, and Michel C. Nussenzweig. Surrogate Light Chain Expressing Human Peripheral B Cells Produce Self-reactive Antibodies. *The Journal of Experimental Medicine*, 199(1):145–150, 2004.

[264] Thomas Tiller, Eric Meffre, Sergey Yurasov, Makoto Tsuiji, Michel C. Nussenzweig, and Hedda Wardemann. Efficient generation of monoclonal antibodies from single human B cells by single cell RT-PCR and expression vector cloning. *Journal of immunological methods*, 329(1-2):112–124, 2008.

[265] Shai Rosenwald, Ran Kafri, and Doron Lancet. Test of a Statistical Model for Molecular Recognition in Biological Repertoires. *Journal of Theoretical Biology*, 216(3):327–336, 2002.

[266] Amnon Horovitz and Meir Rigbi. Protein-protein interactions: Additivity of the free energies of association of amino acid residues. *Journal of Theoretical Biology*, 116(1):149–159, 1985.

[267] Spencer M. Free and James W. Wilson. A Mathematical Contribution to Structure-Activity Studies. *Journal of Medicinal Chemistry*, 7(4):395–399, 1964.

[268] Khouloud A Alkhamis and Dale Eric Wurster. Prediction of adsorption from multicomponent solutions by activated carbon using single-solute parameters. Part II–Proposed equation. *AAPS PharmSciTech*, 3(3):E23, 2002.

[269] Anne-Laure Boulesteix and Korbinian Strimmer. Partial least squares: a versatile tool for the analysis of high-dimensional genomic data. *Brief Bioinform*, 8(1):32–44, 2007.

[270] Svante Wold, Michael Sjöström, and Lennart Eriksson. PLS-regression: a basic tool of chemometrics. *Chemometrics and Intelligent Laboratory Systems*, 58(2):109–130, 2001.

[271] Björn-Helge Mevik and Ron Wehrens. The pls Package: Principal Component and Partial Least Squares Regression in R. *Journal of Statistical Software*, 18(2):1–24, 2007.

[272] R Development Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria, 2009.

[273] Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer, Berlin, new edition edition, 2007.

[274] Simon O. Haykin. *Neural Networks and Learning Machines*. Prentice Hall, 3 edition, 2008.

[275] Wolfram Stacklies, Henning Redestig, Matthias Scholz, Dirk Walther, and Joachim Selbig. pcaMethods a bioconductor package providing PCA methods for incomplete data. *Bioinformatics*, 23(9): 1164–1167, 2007.

[276] V. Franc and V. Hlavac. Multi-class support vector machine. In *Pattern Recognition, 2002. Proceedings. 16th International Conference on*, volume 2, pages 236 – 239 vol.2, 2002.

[277] Vladimir N. Vapnik. *Statistical Learning Theory*. Wiley-Interscience, 1 edition, 1998.

[278] Sepp Hochreiter and Klaus Obermayer. Support vector machines for dyadic data. *Neural Computation*, 18(6):1472–1510, 2006.

[279] Wikipedia contributors. Pearson product-moment correlation coefficient, 2012.

[280] Wikipedia contributors. Spearman's rank correlation coefficient, 2012.

[281] Fionn Murtagh. *Correspondence Analysis and Data Coding with Java and R*. Chapman & Hall/CRC, 1 edition, 2005.

[282] David W. Mount. *Bioinformatics: Second Ed (P): Sequence and Genome Analysis*. Cold Spring Harbor Laboratory Press, 2004.

[283] Gregory R. Warnes. Includes R source code and/or documentation contributed by: Ben Bolker, Lodewijk Bonebakker, Robert Gentleman, Wolfgang Huber Andy Liaw, Thomas Lumley, Martin Maechler, Arni Magnusson, Steffen Moeller, Marc Schwartz, and Bill Venables. *gplots: Various R programming tools for plotting data*, 2011.

[284] Rob J. De Boer and Alan S. Perelson. T Cell Repertoires and Competitive Exclusion. *Journal of Theoretical Biology*, 169(4):375–390, 1994.

[285] Alan S. Perelson. Immune Network Theory. *Immunological Reviews*, 110(1):5–36, 1989.

[286] B Sulzer and A S Perelson. Equilibrium binding of multivalent ligands to cells: effects of cell and receptor density. *Mathematical Biosciences*, 135(2):147–185, 1996.

[287] Sebastian Rausch, Jochen Huehn, Dennis Kirchhoff, Justyna Rzepecka, Corinna Schnoeller, Smitha Pillai, Christoph Loddenkemper, Alexander Scheffold, Alf Hamann, Richard Lucius, and Susanne Hartmann. Functional Analysis of Effector and Regulatory T Cells in a Parasitic Nematode Infection. *Infect. Immun.*, 76(5):1908–1919, 2008.

[288] A.S. Kolaskar and Prasad C. Tongaonkar. A semi-empirical method for prediction of antigenic determinants on protein antigens. *FEBS Letters*, 276(1-2):172–174, 1990.

[289] P Y Chou and G D Fasman. Prediction of the secondary structure of proteins from their amino acid sequence. *Advances in Enzymology and Related Areas of Molecular Biology*, 47:45–148, 1978.

[290] R. Benner, A. van Oudenaren, M. Björklund, F. Ivars, and D. Holmberg. [']Background' immunoglobulin production: measurement, biological significance and regulation. *Immunology Today*, 3(9):243–249, 1982.

[291] N. K Jerne. The natural-selection theory of antibody formation. *Proceedings of the National Academy of Sciences of the United States of America*, 41(11):849, 1955.

[292] M. F. Bachmann, U. Kalinke, A. Althage, G. Freer, C. Burkhart, H.-P. Roost, M. Aguet, H. Hengartner, and R. M. Zinkernagel. The Role of Antibody Concentration and Avidity in Antiviral Protection. *Science*, 276(5321):2024 –2027, 1997.

[293] Tapan Mehta, Murat Tanik, and David B Allison. Towards sound epistemological foundations of statistical methods for high-dimensional biology. *Nat Genet*, 36(9):943–947, 2004.

[294] Sai T Reddy and George Georgiou. Systems analysis of adaptive immunity by utilization of high-throughput technologies. *Current Opinion in Biotechnology*, 22(4):584–589, 2011.

[295] Martijn M. VanDuijn, Lennard J. M. Dekker, L. Zeneyedpour, Peter A. E. Sillevis Smitt, and Theo M. Luider. Immune Responses Are Characterized by Specific Shared Immunoglobulin Peptides That Can Be Detected by Proteomic Techniques. *Journal of Biological Chemistry*, 285 (38):29247–29253, 2010.

[296] D Y Loh, A L Bothwell, M E White-Scharf, T Imanishi-Kari, and D Baltimore. Molecular basis of a mouse strain-specific anti-hapten response. *Cell*, 33(1):85–93, 1983.

[297] Peter S. Andersen, Margit Haahr-Hansen, Vincent W. Coljee, Frank R. Hinnerfeldt, Kim Varming, Søren Bregenholt, and John S. Haurum. Extensive restrictions in the VH sequence usage of the human antibody response against the Rhesus D antigen. *Molecular Immunology*, 44(4):412–422, 2007.

[298] Tine Rugh Poulsen, Per-Johan Meijer, Allan Jensen, Lars S. Nielsen, and Peter S. Andersen. Kinetic, Affinity, and Diversity Limits of Human Polyclonal Antibody Responses Against Tetanus Toxoid. *The Journal of Immunology*, 179(6):3841–3850, 2007.

[299] Dominique de Costa, Ingrid Broodman, Martijn M. VanDuijn, Christoph Stingl, Lennard J. M. Dekker, Peter C. Burgers, Henk C. Hoogsteden, Peter A. E. Sillevis Smitt, Rob J. van Klaveren, and Theo M. Luider. Sequencing and Quantifying IgG Fragments and Antigen-Binding Regions by Mass Spectrometry. *J. Proteome Res.*, 9(6):2937–2945, 2010.

[300] Daniela Frölich, Claudia Giesecke, Henrik E. Mei, Karin Reiter, Capucine Daridon, Peter E. Lipsky, and Thomas Dorner. Secondary Immunization Generates Clonally Related Antigen-Specific Plasma Cells and Memory B Cells. *J Immunol*, 185(5):3103–3110, 2010.

[301] Nina Babel, Juliane Fendt, Stoyan Karaivanov, Gantuja Bold, Steffen Arnold, Anett Sefrin, Evelyn Lieske, Martin Hoffzimmer, Mikalai Dziubianau, Nicole Bethke, Christian Meisel, Gerald Grütz, and Petra Reinke. Sustained BK Viruria as an Early Marker for the Development of BKV-Associated Nephropathy: Analysis of 4128 Urine and Serum Samples. *Transplantation*, 88(1):89–95, 2009.

[302] G. Giraudi, I. Rosso, C. Baggiani, and C. Giovannoli. Affinity between immobilised monoclonal and polyclonal antibodies and steroid-enzyme tracers increases sharply at high surface density. *Analytica Chimica Acta*, 381(2–3):133–146, 1999.

[303] Carsten C Mahrenholz, Victor Tapia, Rolf D Stigler, and Rudolf Volkmer. A study to assess the cross-reactivity of cellulose membrane-bound peptides with detection systems: an analysis at the amino acid level. *Journal of Peptide Science: An Official Publication of the European Peptide Society*, 16(6):297–302, 2010.

[304] Rolf M. Zinkernagel. Immunology Taught by Viruses. *Science*, 271(5246):173 –178, 1996.

[305] Holden T. Maecker, J. Philip McCoy, and Robert Nussenblatt. Standardizing immunophenotyping for the Human Immunology Project. *Nature Reviews Immunology*, 2012.

[306] A. J Burnham, J. F MacGregor, and R. Viveros. Latent variable multivariate regression modeling. *Chemometrics and Intelligent Laboratory Systems*, 48(2):167–180, 1999.

[307] Anders Björner and Richard P. Stanley. *A combinatorial miscellany*. L'Enseignement mathématique, 2010.

Selbständigkeitserklärung

Ich erkläre, dass ich die vorliegende Arbeit selbständig und nur unter Verwendung der angegebenen Literatur und Hilfsmittel angefertigt habe.

Ich habe mich anderweitig nicht um einen Doktorgrad beworben und besitze auch keinen entsprechenden Doktorgrad.

Mir ist die dem Verfahren zugrunde liegende Promotionsordnung bekannt.

Berlin, den 31. Juli 2012

_____

Victor Greiff